EDA on Spotify US Top 200 Charts - Winter 2021, Stats 140SL

Victoria Coronado, Dale Hanks, Chris Lee, Matthew Turk, Sophia Yang and Richard Paul Yim

2 March 2021

Dataset

The data that we have scraped in Python3 using webscraping and parsing libraries; we used the Python wrapper for Spotify's API to pull features, and the typical data wrangling libraries commonly used on Python to construct our data. The final dataset in consideration is of songs that made it to the top 200 chart in the United States in January of 2020 (before the pandemic) and January of 2021, with labels corresponding to a "before" and "after" time period with respect to the occurrence of the Covid-19 pandemic.

Features

The dataset includes 765 observations, each with 16 features. The first two features report the name of the track and the artist, and the last is a factor which points to whether the song reached the top 200 before or after the major outbreak of COVID-19 in the United States (January 2020 = "before", January 2021 = "after").

The rest of the features are numeric variables, most of which are scaled between 0.0 and 1.0. For brevity, we detail only a few of the features:

1. *Danceability* represents how appropriate a track is for dancing, which is based on its tempo, beat and rhythmic consistency; *Accousticness* rates whether or not the track is acoustic; and *Speechiness* rates the proportion of spoken words included in the track.
2. *Valence* describes how musically positive a track is. Higher Valence scores (close to 1.0) are correlated with more cheerful songs, while lower scores (close to 0.0) are correlated with sad or angry tracks.
3. *Key*, which takes on all integer values from 0 to 11, represents the musical key of the track (e.g. C, G sharp, E flat), but does not include denotation of whether the key is major or minor.
4. *Mode* was supposed to signify whether the key was major or minor, but an inspection of the values of this feature revealed them to be inconsistent with the actual keys of many of the tracks that are actually known.

Numeric Statistics

We computed various statistics to the numeric features (we note that these variables are between 0 and 1):

feature	mean	sd	median	min	max	range	skew	kurtosis
danceability	0.69	0.14	0.7	0.14	0.97	0.84	-0.5	-0.08
energy	0.6	0.17	0.61	0.03	0.97	0.94	-0.33	0.17
loudness	-6.76	2.68	-6.35	-29.43	-2.11	27.32	-1.99	9.62
speechiness	0.13	0.12	0.08	0.02	0.64	0.61	1.24	0.67
acousticness	0.26	0.26	0.16	0	0.98	0.98	1.05	-0.05
instrumentalness	0.01	0.06	0	0	0.9	0.9	10.67	129.23
liveness	0.17	0.13	0.12	0.04	0.8	0.75	2.36	6.18
valence	0.48	0.22	0.47	0.03	0.98	0.95	0.15	-0.69
tempo	122.33	29.51	122.96	48.72	205.27	156.55	0.28	-0.51
duration_ms	195783.51	41189.1	193200	30133	386907	356774	0.51	1.67

For ordinal variables, many of the variables had super-majorities towards a single value. We mainly note that between classes of "before" and "after" there were 401 and 364 observations, respectively. We were also interested in seeing the correlation of our input variables. In figure 1a, we *demonstrate use* of a correlation plot. In magnitude, it seems that `loudness` and `energy`, `acousticness` and `energy`, are the two most correlated pairs; additionally, valence seems to be fairly correlated with other features. We can observe other correlated pairs of observations by the correlation matrix/heatmap (on the next page).

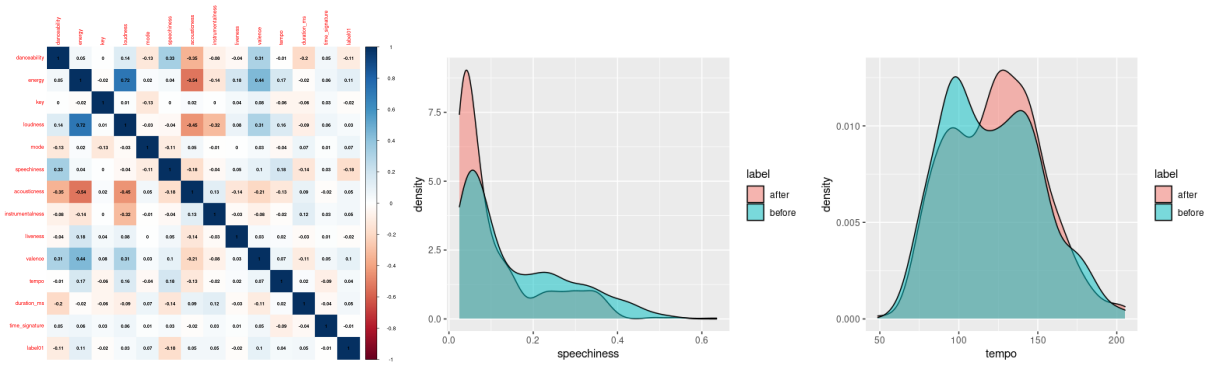

Figure 1: [a, left] We used this correlation map to visualize for ourselves what variables may be highly correlated. [b, center] we layered two histograms of values for each class label corresponding to the speechiness of a song, there is greater speechiness in songs from January 2021. [c, right] Similarly, tempo is higher in January 2021.

Distribution Differences

We were interest in observing some general distribution differences between the population of songs that made it to the top 200 Spotify charts in the US before and after the pandemic, in January 2020 and January 2021, respectively. Referring to figure 1b and 1c, we found that among all the continuously valued variables, histograms corresponding to `speechiness` and `tempo` in tracks seemed to be most differentiated before and after the pandemic (at least visually).

Additionally, we performed non-parametric testing using the Wilcoxon rank sum test to see if there were any statistically significant differences in means of the distributions of our two binary class labels of songs. For an α of 0.05, we found that the following features corresponded to means with statistically significant differences in means between classes: danceability ($p = 0.00168$), energy ($p = 0.00335$), speechiness ($p = 1.082\texttt{e-08}$) and valence ($p = 0.01068$).

Language Processing

Aside from the aforementioned features that we explore so far, we were also curious about some general frequencies of possible track names. Looking at figures 2a and 2b, we have two different word clouds corresponding to the most common words found in a song title by the size of the actual word in the diagram, between January 2020 and January 2021. Generally, we found that "love" is a common word followed by "dance" and "girls" in the entire dataset.

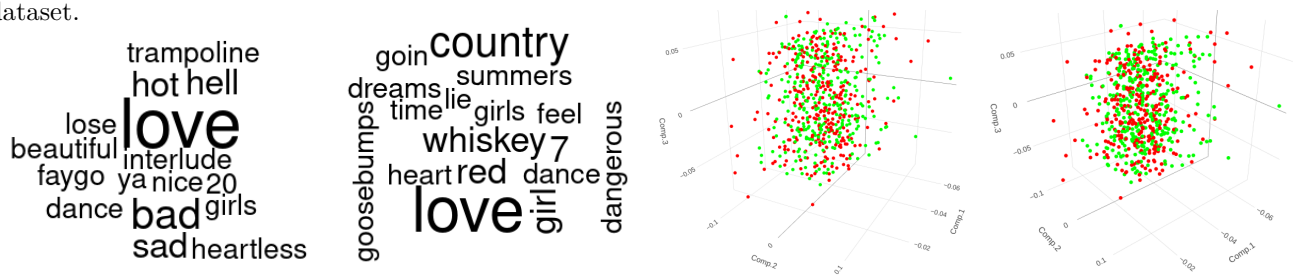

Figure 2: [a, b; left] Word clouds corresponding to songs in January 2020 and January 2021, respectively. [c, d; right] 3D scatterplots with green representing songs in January 2020 and red in January 2021; plots of PCA and SVD, respectively.

Dimension Reduction

We constructed scatterplots on reduced dimensions using principal component analysis (PCA) and singular value decomposition (SVD). Looking at figure 2c and 2d we find that the 3D scatterplots of principal components corresponding to the highest eigenvalues and singular values for PCA and SVD, respectively, show no clear label separation of songs between January 2020 and January 2021. Most distribution differences seem to appear between predominantly in single predictors as shown in our histograms and hypothesis tests.

Future Directions

Based on the above exploration we are mostly interested in discovering what features truly correspond to the most significant variation of songs that made it to the top 200 Charts in the US between 2020 and 2021. We think that we can pursue any, or both, of logistic regression and bagging models to determine these important features.