GIVE ME YOUR PANDEMIC PLAYLIST: A Study of Top 200 Charts Before and After COVID-19

Victoria Coronado, Chris Lee, Dale Hanks, Matt Turk, Sophia Yang & Richard Yim
(Advisor: Vivian Lew vlew@stat.ucla.edu)

UCLA Department of Statistics

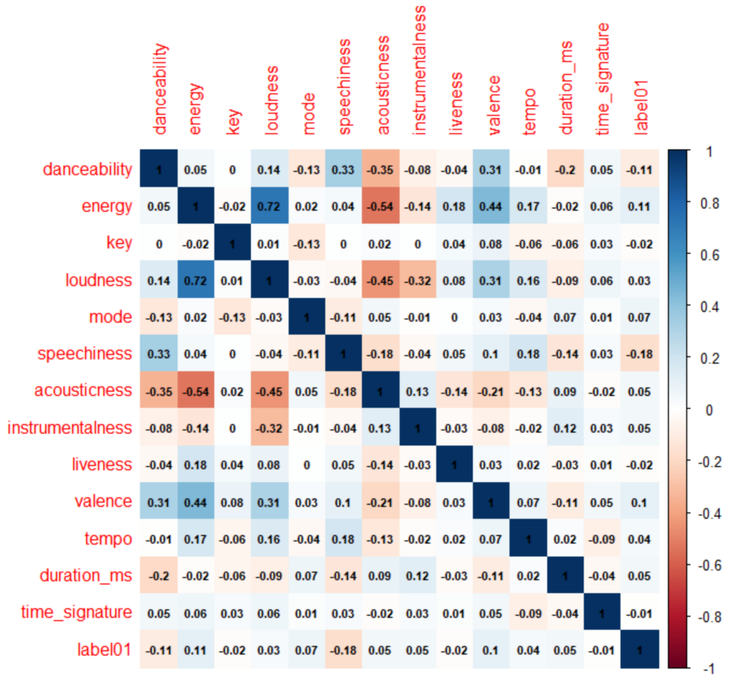
Spotify®

Introduction

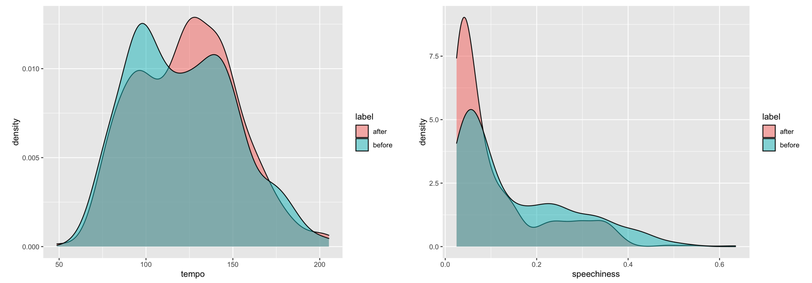
Due to the global outbreak of COVID-19, the year 2020 was unique and challenging (to put it briefly). Struggling with fear of infection, possible job loss and/or relocation, and lack of toilet paper, millions underwent substantial emotional turmoil. As humanity searched for ways to cope with the difficulties we faced, many turned to various forms of entertainment, including music, to pass time at home. The purpose of this study is to examine the possibility of a relationship between the effects of the Coronavirus on popular culture, specifically the top 200 songs on Spotify. Naturally, the hottest tracks will always change with time, but it is still possible to analyze the components of each track, and thereby determine what makes it unique. Using data collected from the Spotify top 200 of January 2020 and January 2021, our team examined 765 observations, each with 16 features, which identify the track, tell whether it made the top 200 before or after the beginning of the COVID-19 outbreak, and rate things like energy, proportion of spoken words, and how much the track sounds like it was recorded live, to name a few.

Features

- **label**: response feature revealing if the song reached the top 200 "before" or "after" COVID-19.
- The first two features correspond to the name of the track and artist.
- The rest are numerical features, most of which are on a scale from 0 to 1.
  - **Acousticness**: rates whether the song is acoustic on a 0 to 1 scale.
  - **Danceability**: how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, and overall regularity.
  - **Energy**: measurement of noise and speed.
  - **Speechiness**: rates the proportion of spoken words included in the track.
  - **Valence**: how musically positive a track is on a scale from 0 to 1. (0=sad/angry 1=cheerful)
- Correlations
  - **Energy** and **loudness** = 0.72
    - As **loudness** increases, **energy** (noise and speed) also increases.
  - **Energy** and **acousticness** = -0.54
    - As **acousticness** increases, **energy** decreases.



Distribution Difference



Of all our variable histograms, **tempo** and **speechiness** show the least amount of overlap. The overall distribution of **tempo** is rather uniform. However, when splitting the data by **label**, **tempo** seems centered around a higher mean for 'after'. The kurtosis for **tempo before** and **tempo after** is -0.5922033 and -0.4030732, respectively. This tells us that the center of the distribution is thicker than expected, or that **tempo** has been consistent to fall within a certain range. The distribution of **speechiness** is heavily skewed left and shows a much higher density of data in the left region for 'after', even with less observations than 'before'. Two possible reasons could be a shift towards music focused more on the instruments and less on voice, e.g. electronic synthetic music void of vocals and following social distancing guidelines in recording studios.

To find out whether the differences in the top 200 Spotify charts before and after the pandemic are statistically significant, we performed hypothesis testing on song features. We noticed from the histograms that features are not normally distributed except **duration**. As they do not satisfy the normality under the Gauss-Markov conditions, we used the Wilcoxon rank-sum test as it does not assume any known distributions of our data, it does not deal with parameters, thus it makes an appropriate test for our data. At a 5 percent significant level, we found the differences in **danceability** ($p = 0.00168$), **energy** (0.00335), **speechiness** (1.082e-08) and **valence** (0.01068) before and after COVID-19 are statistically significant.

Side note: Song Titles

The two wordclouds show that the pandemic has little effect on the most popular song titles, as **"Love"** remains the top 1 word.

Modeling

To determine which features were most instrumental (pun intended) in predicting whether a track hit the top 200 in 2020 or 2021, we began with a logistic regression model. The only three significant predictors are **energy**, **speechiness**, and **valence**. A reduced model was trained with the three predictors. Songs in Spotify's top 200 of January 2021 had higher **energy** and **valence**, while top songs of January 2020 had significantly higher **speechiness**.

Boosting is a general statistical learning approach; we apply boosting using decision trees. Conceptually, boosting iteratively and sequentially grows decision trees by observing the residuals the decisions, in ensemble, during the training process. We implement gradient boosting with decision trees using the 'gbm' library in 'R'. We were interested in achieving a ranking of important variables corresponding to whether or not a song made it to the top 200 charts in the U.S., January 2020 versus January 2021. The gradient boosting process on decision trees naturally does not utilize all variables for each decision tree, so we are able to record a variable importance for how much error we accrue when we leave out a certain variable. We achieved the following ranking of variables after 30 simulations of gradient boosting (we list only the the top five variables):

1. **Speechiness**, with an average decrease (0.152 to 0.109) *
2. **Energy**, with an average increase (0.585 to 0.621) *
3. **Valence**, with an average increase (0.464 to 0.507) *
4. **Acousticness**, with an average increase (0.249 to 0.274)
5. **Danceability**, with an average decrease (0.700 to 0.668) *

Discussion

The two models produce similar results for the feature differences between Top 200 charts in January 2020 vs. 2021. The 2021 ones tend to have fewer words, more energy (noise and speed) and more valence (more musically positive). It makes sense for people to appreciate more energetic and musically positive songs during the COVID-19. Still our study has its limitations: we don't have a large dataset with our 765 observations from two months. Also there might be other confounding factors besides the pandemic that are behind this shift in features of top 200 charts.

References

Spotify API. https://github.com/plamere/spotipy

Spotify for Developers, Design Guidelines. https://developer.spotify.com/documentation/general/design-and-branding/

Gareth James et al. *An Introduction to Statistical Learning: with Applications in R*. Bo Plantinga. "What do Spotify's audio features tell us about this year's Eurovision Song Contest?" https://medium.com/@boplantinga/what-do-spotifys-audio-features-tell-us-about-this-year-s-eurovision-song-contest-66ad188e112a

Julia Silge, David Robinson. *Text Mining with R: A Tidy Approach*, Ch. 2, Sentiment analysis with tidy data. https://www.tidytextmining.com/sentiment.html