Spotify - Group 4: Paul Blankley, Nathaniel Stein, Ryan Janssen, Andrew Lund
Final Project - Milestone 3: EDA & Revised Project Statement
Data Science I - Fall 2017
28 November 2017

## Data Description:

After sourcing the 8466 Spotify users who “own” (the playlists are attached to their accounts) a large number of playlists, the spotipy library was employed to collect user playlists, tracks, user followers, and genre data from Spotify’s API, each saved as JSON files. Genre information for each playlist was harder to source and incorporate than one would think. It does not come directly from each playlist, but rather from the artists of each track as a separate query. In its basic form the playlists file contains 8,231 observations across seven features:

| | collab | desc | followers | id | name | num_tracks | user |
|---|---|---|---|---|---|---|---|
| **0** | False | None | 2.0 | 3ftsSOkyCsILZeAZYQr2jH | Allgott o villgott | 6 | ellenholstad |
| **1** | False | None | 0.0 | 27NLrsj0rlUi9S9Buj7NEl | Emelie och Nelli | 9 | ellenholstad |
| **10** | False | None | 9.0 | 2DVuNt17JxlUUwP8VbjYMZ | Dame mas chocolinas | 54 | maka_97 |
| **100** | False | None | 22.0 | 07nCTAAPUQl3O9835StlnA | Miami Morty 💎🍭💸💘 | 46 | thefamousnobody |
| **1000** | False | None | 0.0 | 2nLFBeJkALxMcmYHqOnBAE | Michael Bublé – Call Me Irresponsible | 13 | vimmel76 |

The tracks file contains 367,949 observations across eight features:

| | added_at | artist | duration | explicit | id | name | playlist_id | popularity |
|---|---|---|---|---|---|---|---|---|
| **0** | 2013-09-30 16:12:24 | Allgott & Villgott | 119907.0 | 0.0 | 3alptaHMnblXRxPWKlqwc6 | Klappa lamm | 3ftsSOkyCsILZeAZYQr2jH | 2.0 |
| **1** | 2013-10-05 15:22:13 | Allgott & Villgott | 67918.0 | 0.0 | 0rPBlDWP6wcfax63Vs8nAF | Hej på dej | 3ftsSOkyCsILZeAZYQr2jH | 4.0 |
| **10** | 2014-07-16 13:50:24 | J Boog | 217270.0 | 0.0 | 4RjHalDdUreXDJSJLo44lK | Sunshine Girl | 35XFuuqgCvTYQARix7CFpm | 53.0 |
| **100** | 2014-07-06 09:12:01 | Brennan Heart | 222919.0 | 0.0 | 6A04TZRVZw8db1VsHeYOEx | Never Break Me - Toneshifterz Remix | 0gGfciue2ZDCOG5uMv46gU | 0.0 |
| **1000** | 2012-06-29 11:53:52 | Johnny Ray | 153375.0 | 0.0 | 1k4p7c69Dkh2b7s813ooR8 | Yes Tonight, Josephine | 6Tuex6CIDfZRyRtsmb5rwE | 0.0 |

And the followers and genre files contain 8,457 and 39,998 observations as follows:

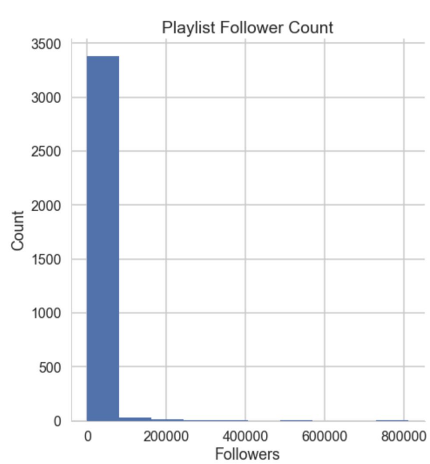
| | user | user_followers |
|---|---|---|
| **0** | 11132487979 | 59.0 |
| **1** | 1214248943 | 29.0 |
| **10** | 1231537904 | 23.0 |
| **100** | asrais | 306.0 |
| **1000** | colib21 | 9.0 |

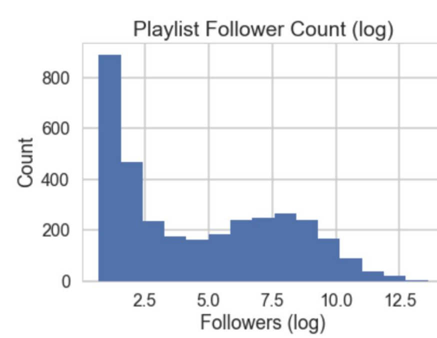
| | artist | artist_followers | artist_genre | artist_id | artist_pop |
|---|---|---|---|---|---|
| **0** | Allgott & Villgott | 76.0 | None | 5psFWO0ApFkCgzjuToEHx3 | 8.0 |
| **1** | J Boog | 103793.0 | [polynesian pop] | 7oEWmZ9dKlAVxTgmjUbYr4 | 63.0 |
| **10** | Trouble Maker | 57509.0 | [dance pop, k-pop] | 0ztjVBmFk6OuHq6XBBwMI9 | 48.0 |
| **100** | Whigfield | 14986.0 | [bubblegum dance, dance pop, eurodance, europo... | 0lHoDF96DNKSIclpcOfMnq | 56.0 |
| **100000** | Los Violadores | 15328.0 | [argentine rock, latin alternative, latin meta... | 4EkrhlCS2DbFxvC3Uhq6p2 | 45.0 |

The files were combined and manipulated into a single DataFrame via the playlist ID, username, and artist name attributes. While cleaning, we decided to eliminate any playlist with only one follower (the owner).

## EDA:

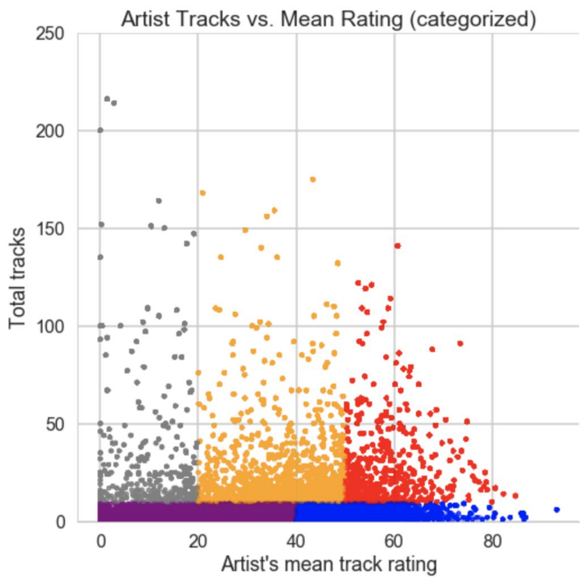
Engineering more features around artist and playlist data yielded interesting relationships. A cursory measure of playlist success, our generalized project goal, is the number of followers per playlist. The histograms to the right show a significant right-skew, with most playlists having few followers and a few approaching the 1M mark. The log-transform of followers illustrates a potentially more useful response variable.

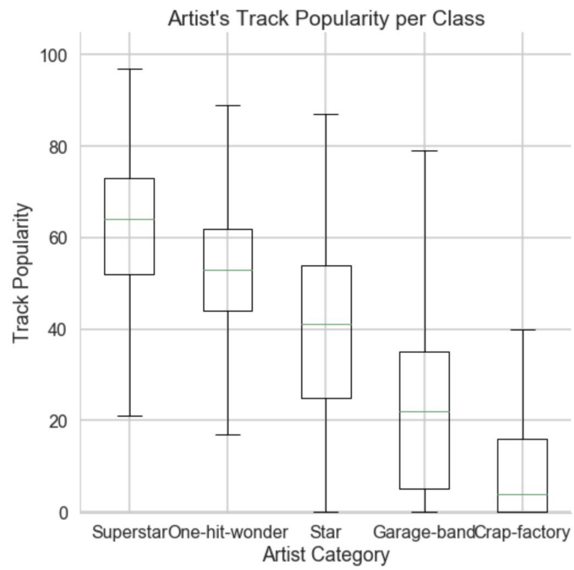


We separated artists into five distinct categories using combinations of their mean popularity and total number of tracks as follows: **superstar**, **star**, **one_hit_wonder**, **garage_band**, and **trash_factory**.

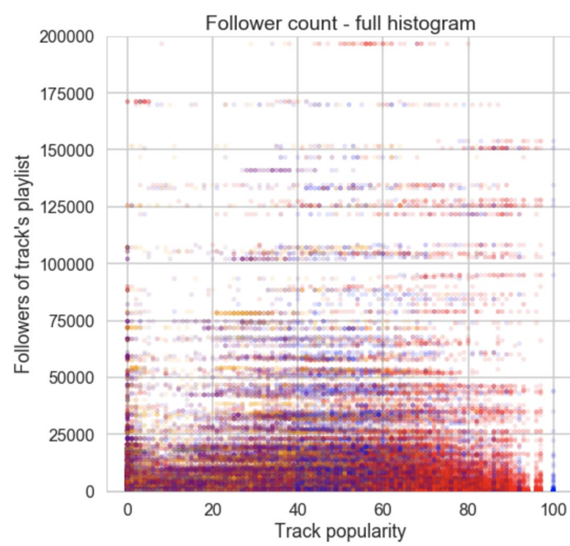
These categories should be self-descriptive, but the two plots below do a fantastic job of illustrating each in terms of their thresholds and relationships. The artist popularity metric ties out with our expectations because we see people like Post Malone, Camila Cabello, and Ed Sheeran in our superstar category.

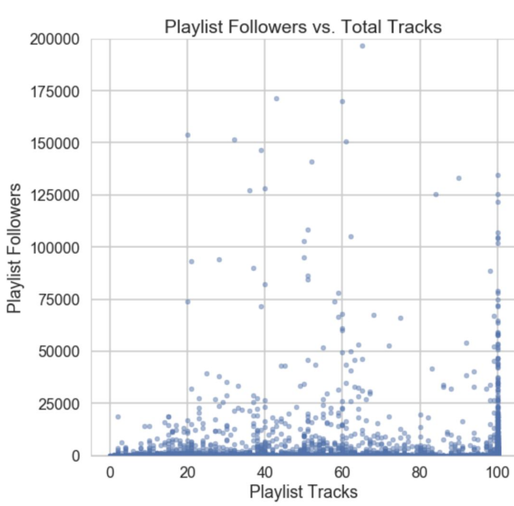


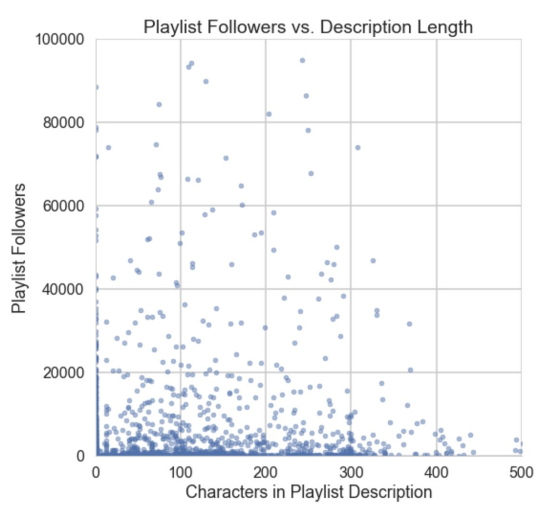




Comparing these five artist categories with specific track popularities (as determined by Spotify) and those tracks' playlists' followers shows us that not only do most playlists incorporate popular artists (red/orange), but a lot of them have the "superstar" artists' best songs (the mostly red right side), as well as a smattering of "one-hit-wonders" (blue in the middle) and a dense area of moderately popular songs throughout (between 30 and 70 popularity). This relationship and playlist architecture should also be intuitive, and the scatterplot to the left illustrates it well.

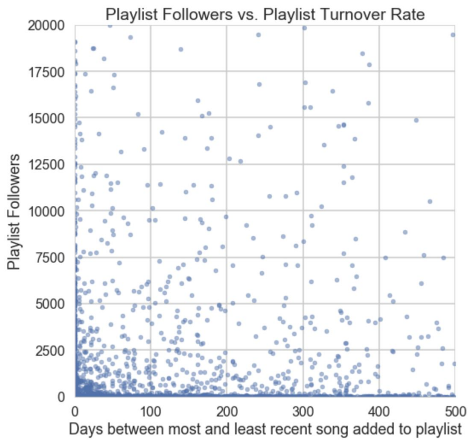
It appears that there may be a relationship between the number of tracks a playlist has and its number of followers (below), as well as length of its description (below-right).

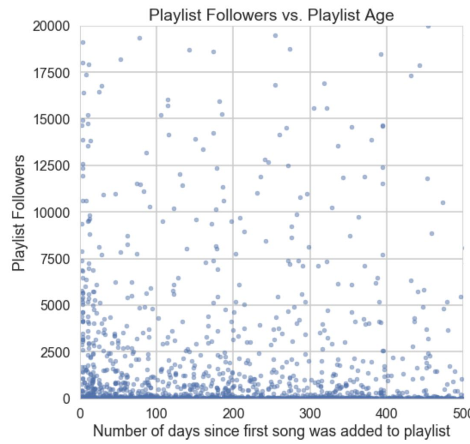


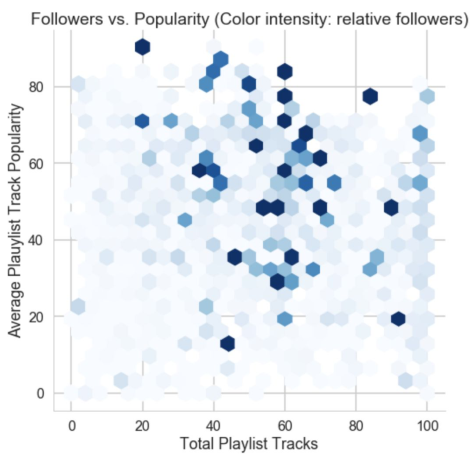



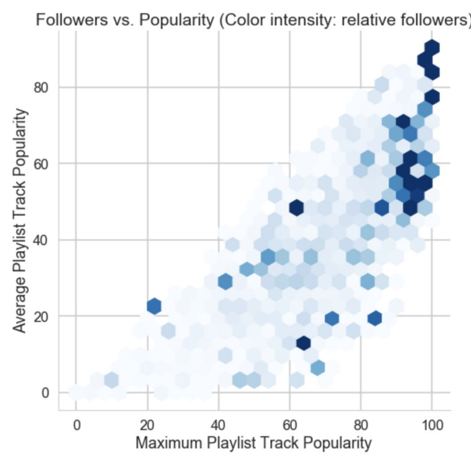
We also investigated the impact of playlist turnover rate, or relative age (days between the oldest and newest track being added), as well as the the absolute age (days since first song was added) on playlist follower count (below plots).




Two more illustrative hexagon bin plots are below. They show interesting relationships between relative number of playlist followers with playlist popularity and track metrics.

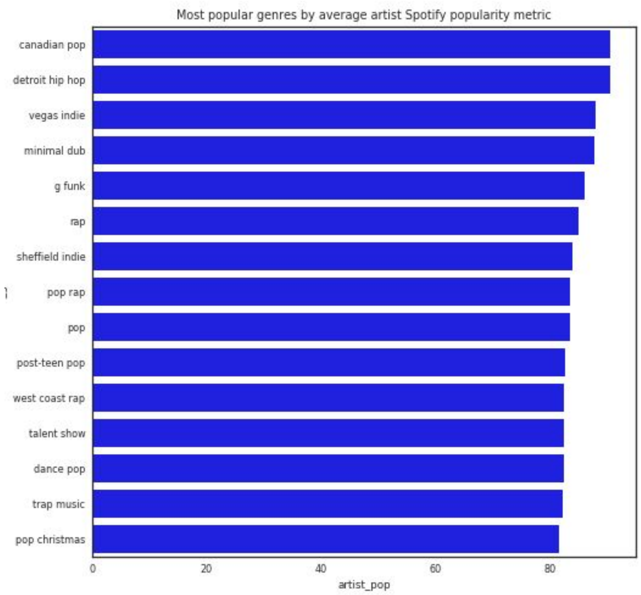


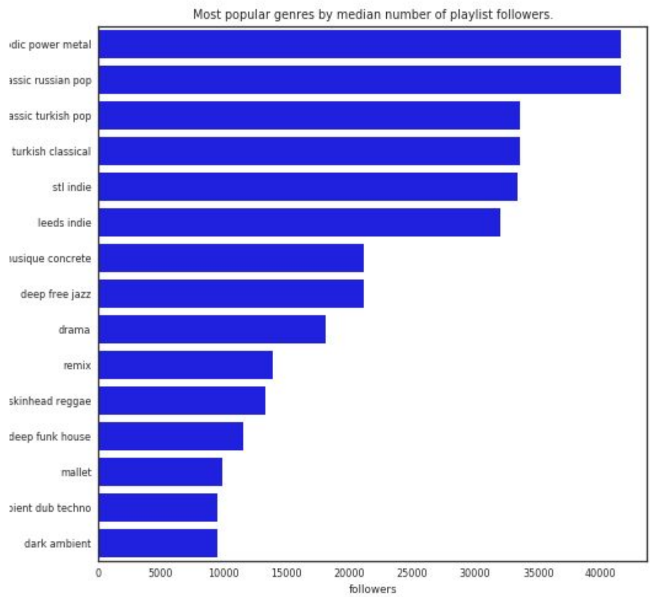




The overarching goal of the preceding EDA was to find trends in the playlist data indicating “success” of a playlist. How do we measure that success? The simple metric of total followers may not the best measure. In that regard, we propose use of a response variable analogous to “Playlist Velocity” (PV), or playlist followers over time. The time component may change (time since inception, time since last song added, mean age of song in playlist, etc.) during model analysis, but PV will remain as our response variable “theme.”

We also plan to incorporate track audio data from the Million Song Dataset (MSD), and hope to combine some of those features into a significant composite predictor for PV. One downside to including this audio data is that it may shrink our overall dataset since some of the songs included in playlists may or may not be included in the MSD. There will be more to follow on that front in the final submission report.

In the below graphs, we look at the influence of genre on popularity.  On the left we can see the genres associated with the highest mean popularity for the artist.  These results make sense because “canadian pop” can likely correspond to Justin Bieber, and “detroit hip hop” can likely correspond to Eminem, and so on.  The graph on the right is a little more thought provoking.  We see some rather strange genres with the highest median playlist followers.  I think this implies that these genres show up very few times, but when they do, they often show up in very popular playlists.




## Revised Project Question:

Our team's original two project questions were intentionally general. The preceding EDA has cemented their applicability to this project: Can we predict the general success of a Spotify playlist using regression? Can we classify a playlist into different popularity classes using the predictors we use for the regression model?

Based on our data collection and exploration, we have revised our questions by maintaining the original two and adding a third: Can we generate a successful Spotify playlist with user specified genre and length using our regression model?