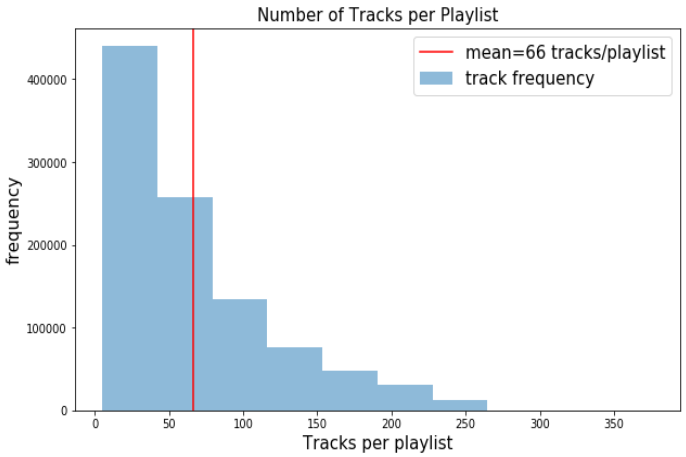
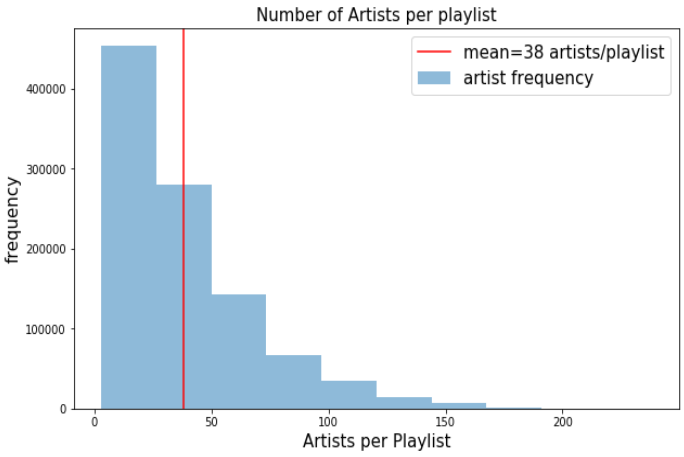
CS 209a Final Project: Spotify
Milestone #3
11/28/2018

Research Question & Project Objective: In this project we set out to develop and test a model that answers the following question: Can Spotify improve end-of-playlist track recommendation diversity, without sacrificing user engagement? Specifically, our goal was to deliver a reasonably accurate track recommender that predicts the next 10 tracks for the Spotify users' playlists with the inclusion of less known, mid-tier female artists.

Progress & Modifications: While our research question and overall objective remain unchanged, our modeling ambitions have been somewhat dampened by the unruliness of the Spotify data and the time constraint for this project.

- **Data Bottlenecks**: The dimensionality and the disjointed nature of Spotify data have presented data wrangling challenges beyond our original expectations, leading to modeling delays. The team was faced with a steep learning curve when it came to sparse matrices. Even after the data was cleaned and merged into a user friendly format (over three days via Michael Emanuel's deep SQL expertise), we were left with a data set that contained 5-367 features (columns) for each playlist. Overall,we have 66 million playlist entries. We will use those playlists with at least 20 entries (since we are forecasting the last 10), which takes us down to 64 million entries. Given that dimensionality reduction tools like PCA would not work in this case, we were left with impractically long training times for our model. For example, predicted run time on one of our models was 106 hours. Thankfully, we were able to leverage Michael Emanuel's SQL expertise in database building and modeling, which sped up our model run time significantly (down to 5 hours).





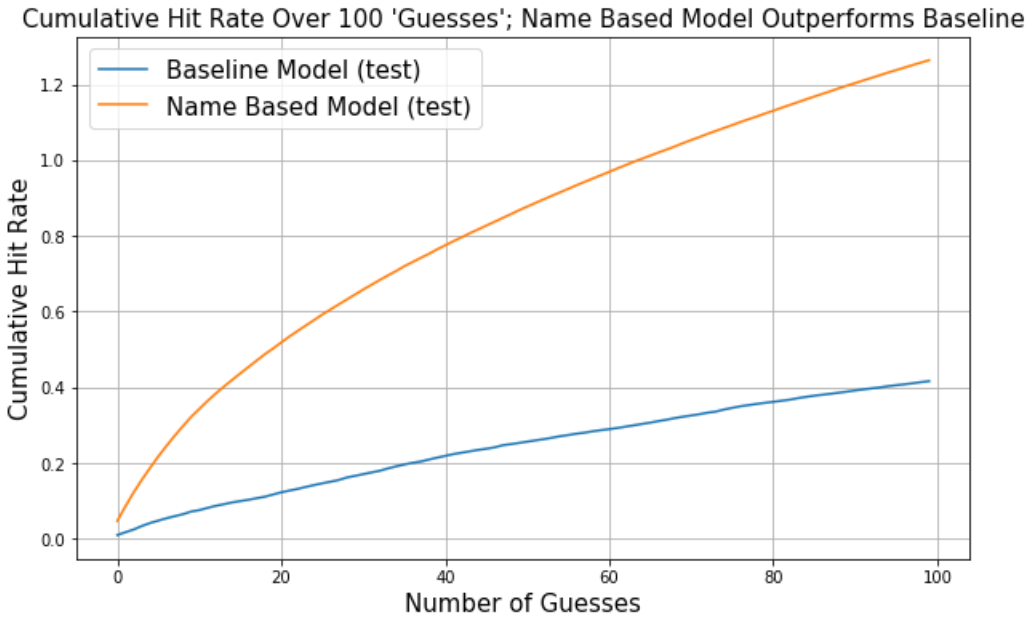
- **Model Modifications**: We made 3 key modifications to our modeling approach since Milestone 2:

    - **Three Models**: Instead of building 2 models (one simple baseline frequency model and second informed recommender model), we are now looking to build three models:
        - Simple baseline frequency model for track recommendations based on tracks' probability distribution. (i.e. the probability of guessing correct tracks in the last 10, based on overall tracks' frequency adjusting for tracks that were already in the playlist).
        - Playlist name based model that will train a recommender using playlist name and tracks in that playlist as well as tracks in other playlists that share the same name.
        - Track Pair Based Model/Naive Bayes: Here we will attempt to reach beyond the scope of the course and look at the relationship between track pairs in each playlists and use it to forecast the last 10 tracks.

    - **Relative Performance Metrics**: Instead of using cross entropy, we will measure our model performance more simply by looking at how model 2 and model 3 perform on test data vs. our simple baseline model (model 1).

    - **Measure of Overall Success**: We will be sending each of the survey respondents an individualized recommendation containing 10 tracks based on the playlist information collected in our survey. Currently , the average US premium Spotify user streams female artists about 21% of the time. Since our goal is to improve playlist diversity, we plan to include 3-4 tracks from the mid-tier female artists in the list of 10 recommended by our model. Our measure of success will be defined by the survey participants' overall satisfaction with the playlist we have provided. We will be especially interested in their response to the lesser known, female artists' tracks.

- **Preliminary Results**: As of 11/28, Model 1 and Model 2 were built. We scored the models by making 100 guesses for the last 10 tracks in a playlists and then seeing how many out of those 100 were correct. On average our baseline model accurately predicted 0.416 tracks on training data and 0.416 tracks on test data. Our playlist name based mode did significantly better, accurately predicting 1.665 tracks in the training set and 1.264 in our test set. While on stand alone basis, these results do not seem very impressive, on relative basis we now have a model that does 3-4x times better than a moderately educated guess. We believe that this is something to get excited about.



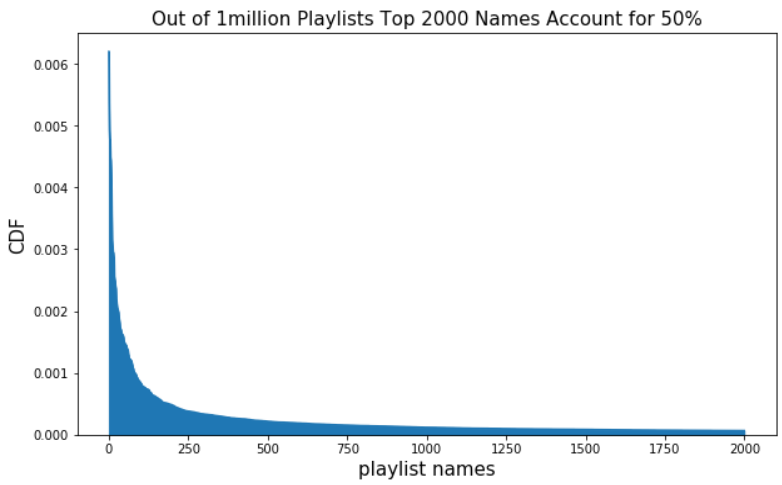
- **Survey Update**: With our playlist survey, we attempted to collect a random sample of music listeners from various social media channels, including Harvard students, work colleagues, college classmates, family and friends. While our response rate was encouraging at 186 participants, after cleaning the survey data we ended up with just 42 complete, usable responses. Our goal is to have track recommendations finished by December 1st and to start following up with the respondents. In the follow-up survey, we will give them an opportunity to rate our recommendations on various aspects.

<u>Important Visualizations</u>:

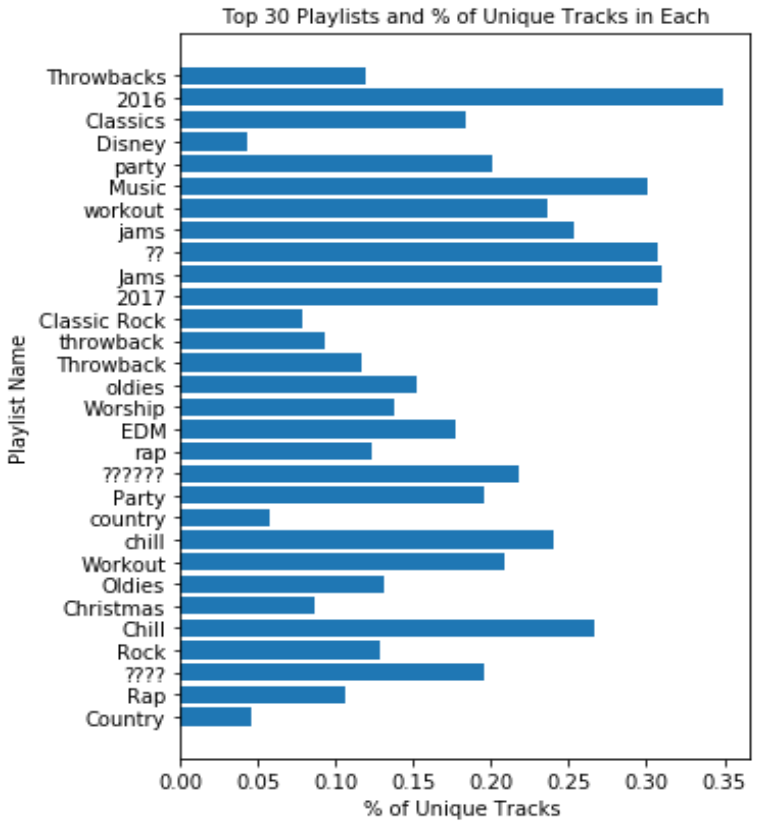
While Spotify offers plenty of opportunity for its users to create unique playlist names, most users end up picking the same playlists names as their peers. In fact, out of 1 million playlist, half are described with just 2000 names (i.e. a lot of repetition).The table below shows top 5 playlist names:

	PlaylistName	name_count
0	????	6197
1	Country	5388
2	Chill	4945
3	Rap	4793
4	Christmas	4779

To make things interesting, 'nameless' playlist is the most popular choice.



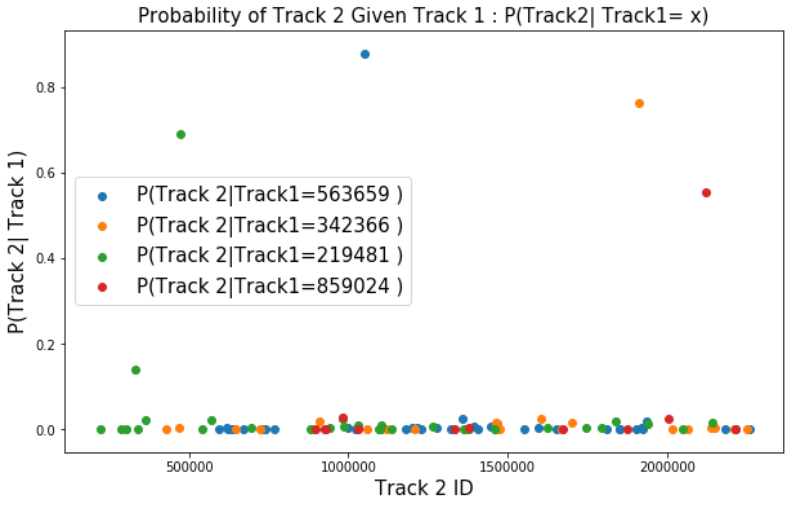
We observe the same trend when it comes to track selection. On average, ~30% of the total track selections per playlist name are unique, with the rest being duplicated in another user's playlist with the same name. This was the key inspiration behind the playlist name model that takes into account these similarities between identically named playlists.
The plot on the right illustrates this observation. In addition, it also illustrates a playlist name issue where we have some of the equally named playlists screening as different due to choice of letter capitalizations and of course we note several 'nameless' playlists When dealing with '????' playlists, we made a decision to combine them into into one simplified empty string name (since they do appear to have some tracks in common). Similarly, we used simplified playlists names that collapsed the same names (albeit differently formatted) into one by stripping out punctuation and capitalization. This takes us down from 74k playlist names to 17k playlist names and helps computation efficiency.



Our track-pairs data set includes 151 million entries. Thus, for the purposes of the chart on the right a sample of 1 million entries was drawn and top frequency track pairs were select as per table below:

	TrackID_1	TrackID_2	Frequency
65722194	563659	1050231	5109
41727599	342366	1911671	4441
27356794	219481	471925	4002
93146032	859024	2120994	3215
82460860	728971	1920170	2812

We note that knowing ID of Track 1, gives us very useful information about the probability of certain tracks (Track 2) being in the same playlist as track 1. We ran four examples here to illustrate this point.



Appendix: Audio Features vs. Track Popularity

While, some interesting relationships could be gleaned from the charts below, given our time constraint we will not be including these features into our model.