Collin Smith

Conor Burrows

Jake Luedtke

Nicholas Melon

CSC 255

Our Goals and How They've Changed

We set out to characterize the popular music of recent decades, and what types of songs were on top of the charts at different periods of time. We were mainly thinking about genre distribution at first, and how that has changed year to year. We also wanted to consider the words commonly used in song titles, and see if certain words yield better scores, or if certain words were used more in particular genres. As we worked with our data, we were slowed by the process of data acquisition and cleaning. Once we had usably formatted data, we found several variables (mood, tempo, artist type) that we hadn't been considering, and our options opened up. As we worked with our data and conceived of new ideas, we put the pieces together to make the product which we present today.

Our Data

The datasets that inspired this project came from tsort.info, as did the scoring system without which much of our work wouldn't have been possible. The score for each song is calculated using a formula which incorporates all chart positions and awards from tsort's sources. With these scores, we can easily compare different songs' popularity and critical reception. The data from tsort is very conveniently formatted--the website developers clearly intended their data to be used for research such as this. However, the information in tsort's charts is limited. We had no description of genre, or any characteristic of the songs, for that matter, besides title, artist name, and year of release. To add genre to our dataset, and make possible all the fun analyses which we had in mind, we turned to Gracenote.

Gracenote Saga

In order to answer questions about the popularity of music, we needed info from a more in-depth source. Scores and z-scores are good, but they don't tell us what gives a song it's "groove." In addition to popularity, we also needed data related to the contents of a song. Gracenote, a music information database, made it very easy for us to fill in things like tempo, genre, and mood for all of the tsort songs. Using "Pygn" and "Gnaw" (the latter being our custom wrapper), we completely gathered metadata for tsort's "Top 5000 Most Popular" songs. During the initial collection, we discovered that Gracenote also included fields such as "artist type" that lead to more interesting analysis. Paired with the initial list, Gracenote helped us find out more about what makes a song popular as well as what aspects popular music shared.

Answering Questions with Visualizations

The first question we thought about was "how have popular genres changed throughout the years?" The segmented bar chart below (see figure 1) helps answer this question. Each color represents a different genre, and the size of each segment represents the proportion of that decade's songs that were of that particular genre. The genre which is almost always prevalent is pop (dark red colored segments). It seems to dominate the charts, along with one or two other genres which are different depending on the era. For example, in the 1920s, 30s, and 40s, pop shares the stage with jazz music, whereas toward the 2000s, Rock and Urban music are larger competitors. The exceptions to pop's dominance are the 1960s and 1970s, when Rock songs made up a much larger proportion of the top songs. However, Rock's popularity became less widespread toward the 2000s, with pop songs becoming dominant again, and the growth in popularity of the Urban genre.

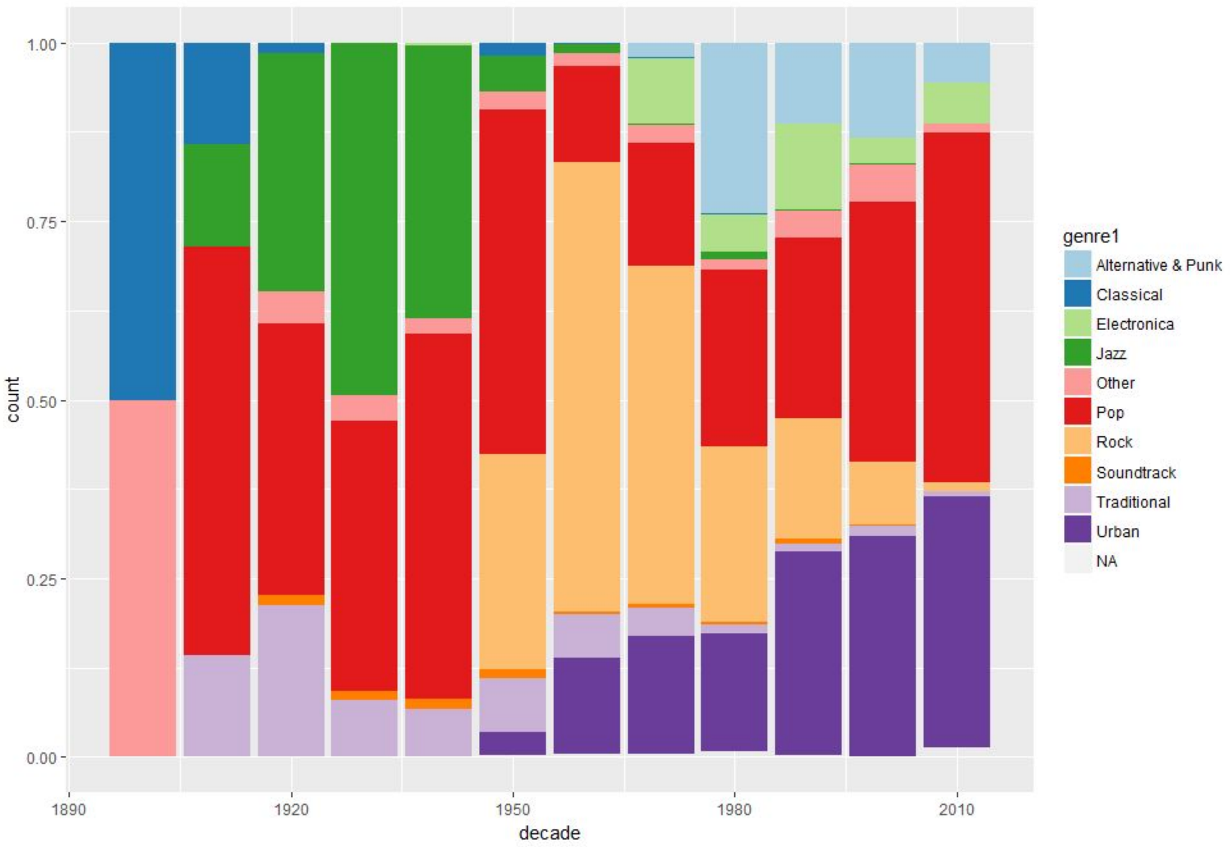

Figure 1 ↑ Percentages of Genres by Decade

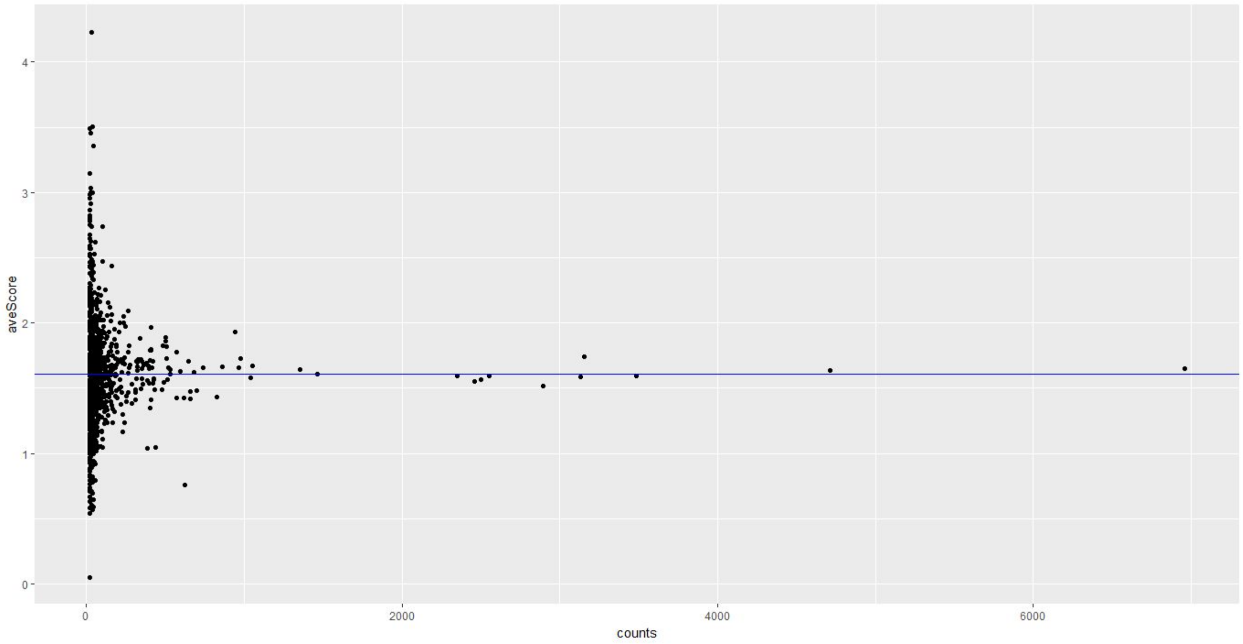

Figure 2 ↑ Title Word Count vs Average Score

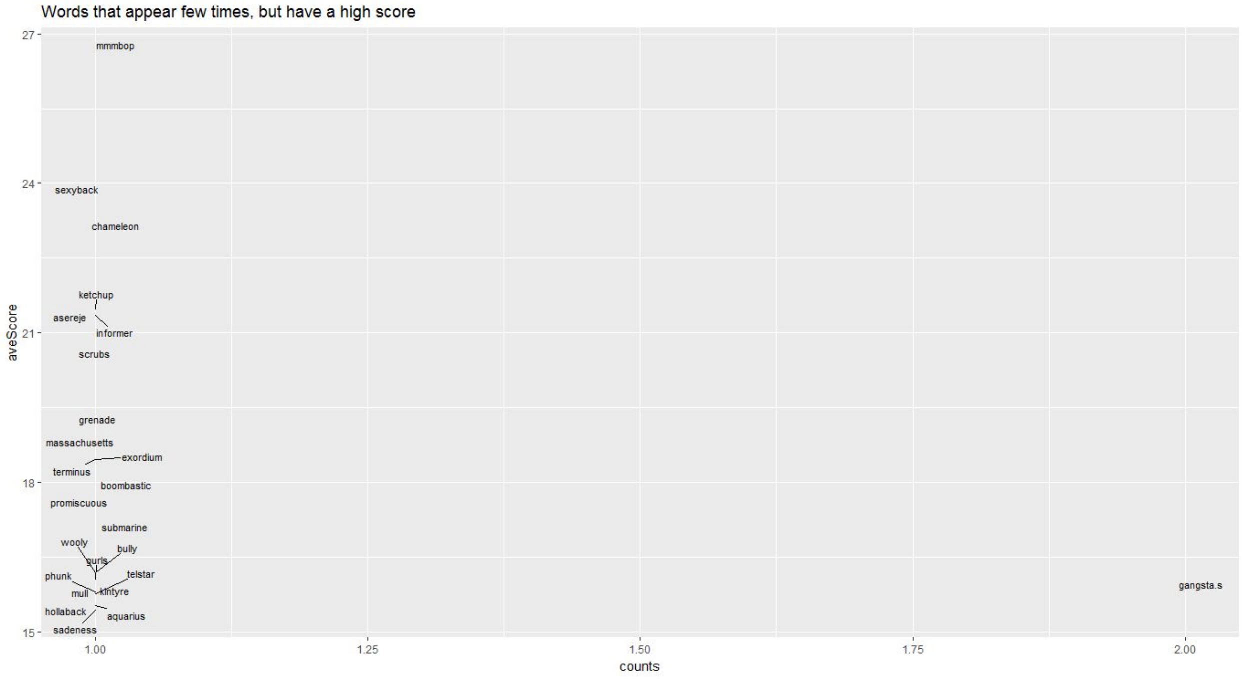

Figure 3 ↑ Low-Count Title Words with High Scores

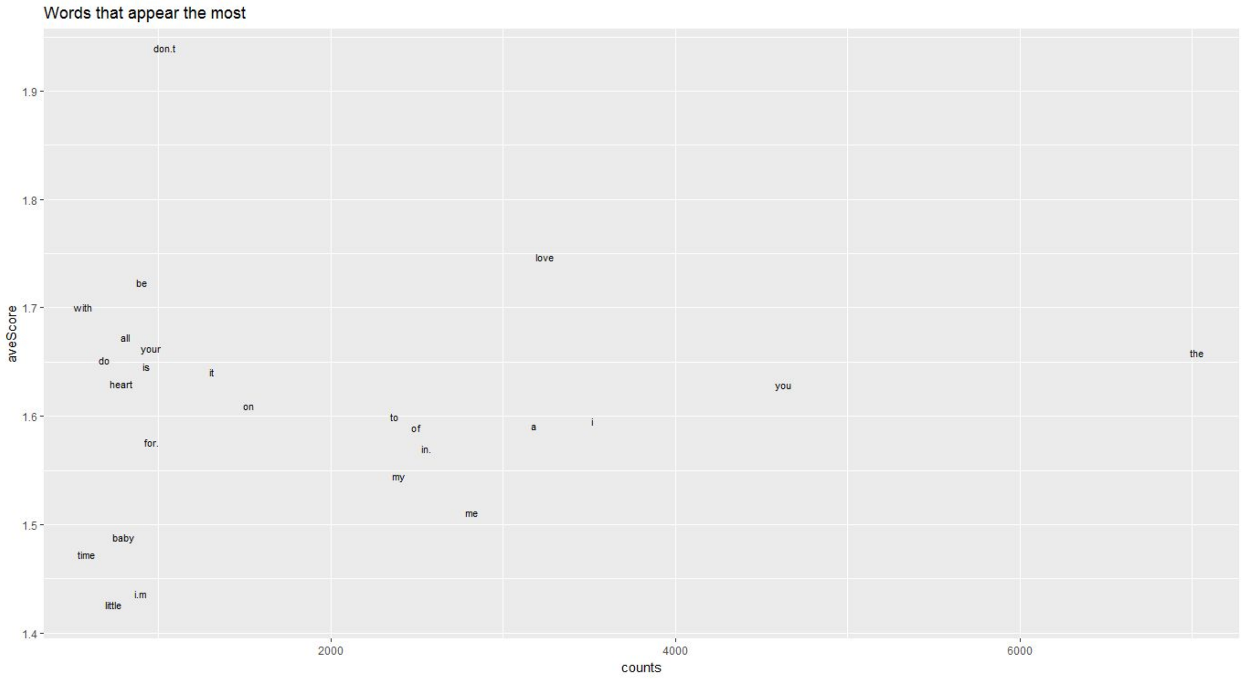

Figure 4 ↑ Most Frequent Title Words

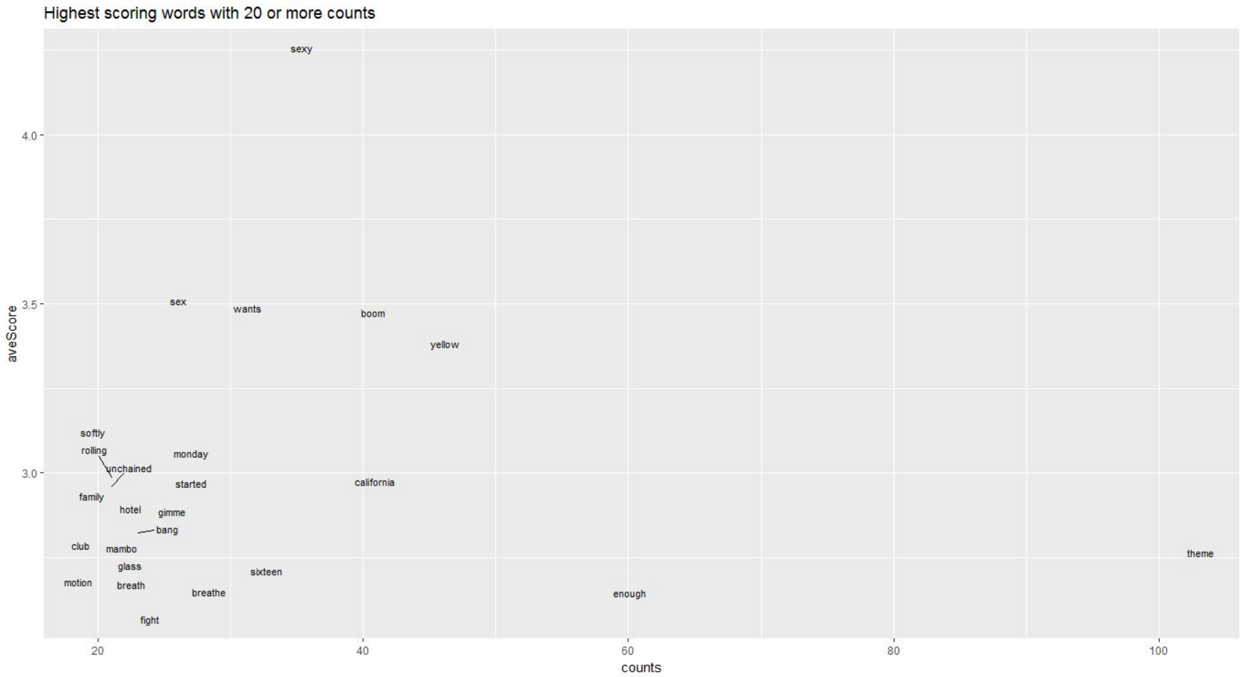

Figure 5 ↑ High-Scoring, High-Frequency Title Words

Word Analysis

In our goals we mentioned that one area of interest was if the words in a song title had any relevance to its popularity or the type of music it's in. To answer problem we looked at our 48000 songs database. We made a list of all the words that appear in the dataset and then when a repeat word appeared we added plus one to its count. With this dataset we made some scatterplots with an average line at one(figure 2) and we noticed a lot of words centered around it. To see what words they were we made another graph(figure 3) and saw that these words were very common names. Next we wanted to see what words were the most popular and to do this we charted the first twenty-five words and sorted them by average score(figure 4). This graph showed us that most popular words were the ones that appeared once or twice and not in the graph these words were also the ones that were the least popular. It makes sense that this is the case, due to the fact that if there is only one or two observations that they won't represent an average. To get meaningful data on a word's popularity we decided we wanted to look at words that appeared more than twenty times so it could represent an average. We used this data to make figure 5 and from it we see some interesting results. For example we can see that words like sexy and sex make for popular songs. Reasons behind this could be that these words attract a person’s attention and same with words like boom and bang. A point on this graph that is different from the rest is the word theme. This word is popular due to the fact that some songs in the database where from soundtracks in movies like Star Wars and they were listed as “Theme from a Movie example”. After we finished analyzing the words in the our forty-eight thousand song database our meta data was done for our top five thousand.

Once our meta data was appended to our top five thousand songs with the z-scores, we wanted to see if a word’s definition had any effect on its average genre and mood. To do this we got the count of every word in this database, got the location where it appeared, and took the average of the genres, moods, and other columns. After this finished we began to review our results and saw that it was as expected. Words like love and alone , had an average mood of dramatic / emotion and average tempo of slow, which could be expected for words like those. It was a nice end result though to see if a word in fact had an effect on a song’s attributes.

Predictions

We wanted to attempt to create prediction models that could predict attributes of songs based on basic input about that song. Although we were not able to predict score like we initially planned, we were able to predict the genre of a song with inputs such as year, tempo, and score. When the model was tested on a test set of 899 songs, the model predicted the genre correctly on average 402 times out of 899 songs. Although this is less than fifty percent we believe that this is partially because some songs were given multiple genres and others were broken down into small sub genres. This also meant that there was a clear correlation between our predictive variables and the prediction variable. To attempt to see what variables are correlated with genre we removed one variable at a time and observed its behavior. To our surprise when tempo was removed, the predictive model seemed to perform almost the exact same. We concluded that the tempo of the song was not correlated with genre. However when either score or year was removed the model performed significantly worse. These predictive models have lead us to believe that there is a correlation between score, genre and year that we would like to further explore in future research.

Rise of Pop and Urban Post-1990

You might notice that in the 1990s and forward the “urban” and “pop” genres seem to take off even more than before, and after some digging we believe that this could be related to the 1996 Telecommunications Act. Among other things, the Act allowed companies like Cumulus to purchase multiple radio stations and control what they played. Eventually this lead to a standardization of broadcasting, and over the next five years almost every radio station in the country changed formats. Part of the change included radio giants taking music directly from labels, and as the recording industry changed it became more profitable to promote premade music instead of scouting for new bands and talents. The “rock” genre suffers even more after the 90s, whereas pop music takes off due to the market being saturated with solo and duo artists.

Looking Forward

There are countless ways to expand on this research. Primarily, we’d like to work with larger datasets. The dataset we used for most of our analyses was just the 5000 best scored songs since 1900, which may not be representative of music in general. With more time, we could use the Gracenote API to gather metadata on more songs. Then, we could perhaps make farther-reaching conclusions, and be more confident about those conclusions.

Time also limited us in terms of the relationships we were able to explore. With so many interesting descriptors for each song, we had a hard time deciding what questions we should try to answer. In further research, we might ask whether certain moods or genres have faster tempos. We might further explore the differences between songs by female and male artists--what moods are preferred by each; is there a difference in tempo; are certain words used more by one type of artist. Additionally, we could explore the many possible categorizations of genre. Gracenote provides three different genre labels, each with a different level of specificity. We chose the most general, since more specificity made some groups unreasonably small. However, using the subcategories, we might ask how each umbrella genre (Pop, Rock, Urban) is split up, and how diverse the songs of a single genre can be.

These are only some of the questions we might explore with more time. With this project, we had to pick our battles. We're happy with the analyses we performed, but the possibilities for future analysis are endless.