Team 6A: Salem Arthur, Marianna Carini, Eric Chen, Edward Shih-Yu Chung, Haiying Ge

**SPOTIFY NETWORK ANALYSIS**

**Introduction**

Since launching in 2008, Spotify has grown to become a global leader in the digital music, podcast and video streaming space by providing access to millions of content to users across the globe. The Swedish company with its “Freemium” pricing model and strategy has been able to gain considerable market share from its competitors like Apple Music, Tidal, Deezer and Amazon Music. This was further bolstered by a strong 2020 Q4 report; outlining a whopping 155 million paid subscribers while growing its total user base by 27%.

Considering the keen interest that our team shares in the entertainment industry, we settled on utilizing Spotify’s extensive API after considering the insights that we could garner from understanding the various relationships that possibly existed between the different features of music being streamed and their impacts on the relative popularity of a song. Streaming companies as a whole reported a combined $10 billion in revenue for 2020 and for the music industry in particular, it accounts for 83% of the revenue made. Due to the importance of streaming services, it would be worthwhile for record companies and other stakeholders, to understand the various variables that contribute to the popularity of a particular song.

Our dataset contains over 175,000 songs collected from the Spotify Web API for music from 1920 to 2020. Thus, we will leverage the vast data to find important relationships between different artists from our analysis. This can provide significant insights that could help us better come up with a prediction on which artists/music will dominate in the US on Spotify.

**Feature Definitions**

1. ***artists***: The list of artists of the song.

2. ***danceability***: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.

3. ***duration\_ms***: The duration of the track in milliseconds.

4. **energy:** Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy.

5. ***explicit***: Binary field identifying whether a track is explicit (1) or not (0).

7. ***instrumentalness***: Predicts whether a track contains no vocals. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content.

8. ***key***: The key the track is in. Integers map to pitches using standard Pitch Class notation . E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on.

9. **liveness:** Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live.

10. ***loudness***: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks.

11. ***mode***: Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.

12. ***name***: Name of the song.

13. **popularity**: The popularity of the track in the US. The value will be between 0 and 100, with 100 being the most popular. The popularity of a track is a value between 0 and 100, with 100 being the most popular. The popularity is calculated by a proprietary Spotify algorithm.

14. ***release\_date***: The date the album was first released, for example “1981-12-15”.

15. ***speechiness***: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audiobook, poetry), the closer to 1.0.

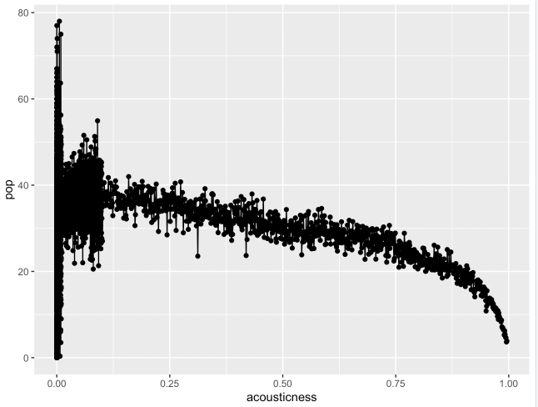
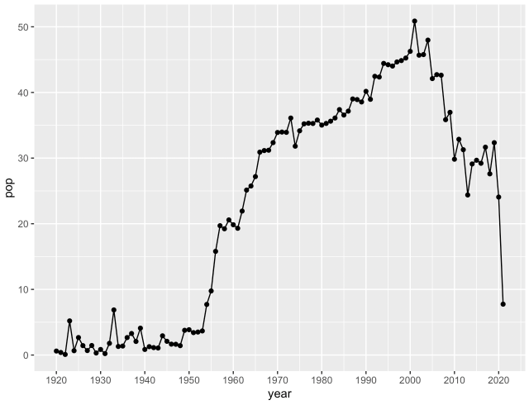
16. ***tempo***: The overall estimated tempo of a track in beats per minute (BPM).

17. ***valence***: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

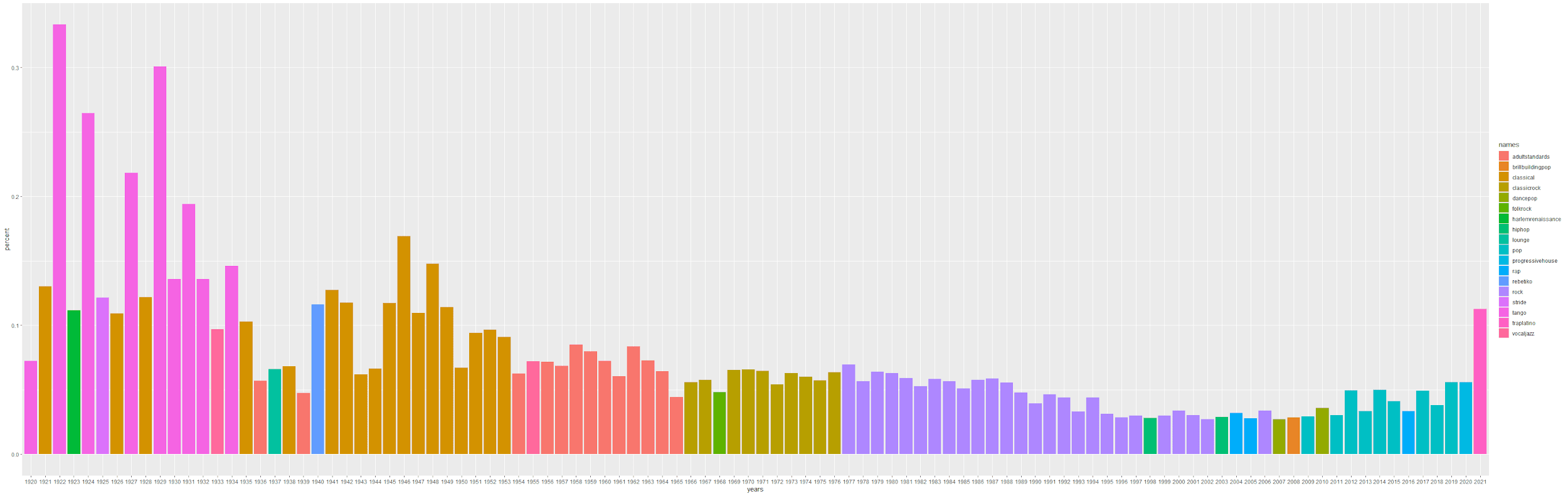
18. ***year***: Year information extracted from release\_date.

19. ***genres***: A list of the genres used to classify the album. For example: “Prog Rock”, “Post-Grunge”. (If not yet classified, the array is empty.)

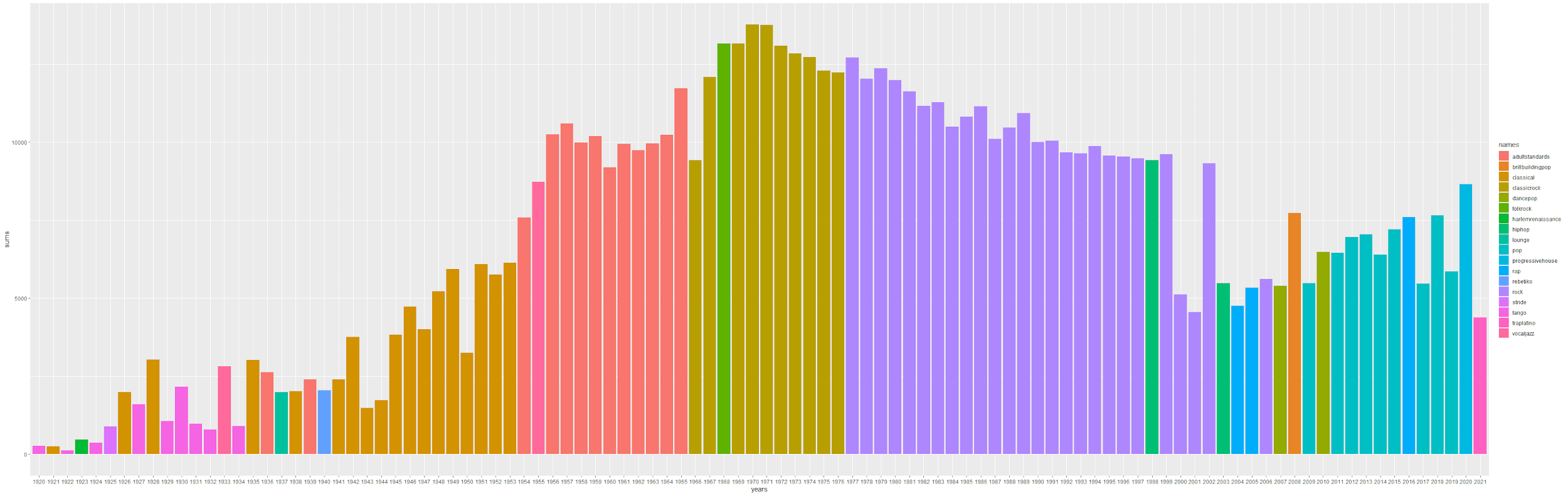
**Exploratory Data Analysis**



The above graph on the left shows the mean popularity of songs by year, indicating that the most popular music tends to be music from 1960 to 2010. This graph on the right above indicates the song’s popularity vs. acousticness. Based on the graph, the popularity drops along with the acousticness score, meaning that people tend to like music gone through electrical amplification.



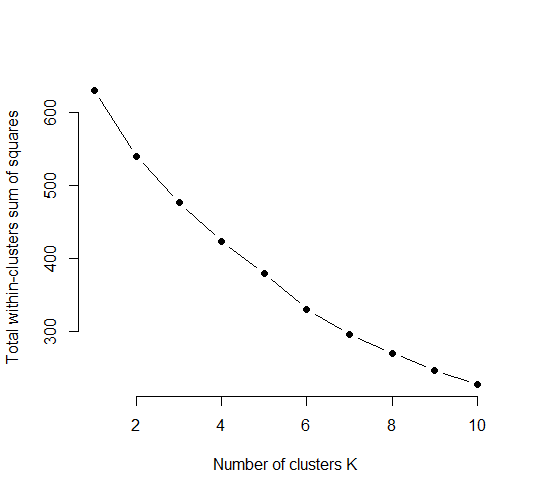
The above graph shows the most popular genre each year as a percentage. There are two major findings here. Firstly, most popular genres are clustered in eras. In other words, there were genres that each dominated for multiple years. Secondly, in the 1920s and 1930s, most popular genres were more dominant than in recent history. One potential explanation of this is we have more diversity in music over time likely due to better technology and the internet.



Similar to the previous graph, this graph shows the most popular genre each year with the total number of genres in a given year. The key finding here is that music with the most variety of genres is from the 1960s to 1990s. Contrary to our previous assumption, the variety of music was not growing through time. The potential explanation is that people during those times didn’t have as much entertainment as we do nowadays, and the music was a main source of entertainment for them. Thus, people in that era tried to make various kinds of music.

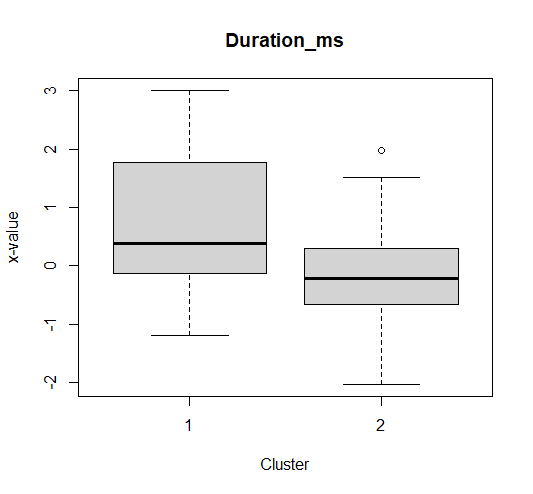
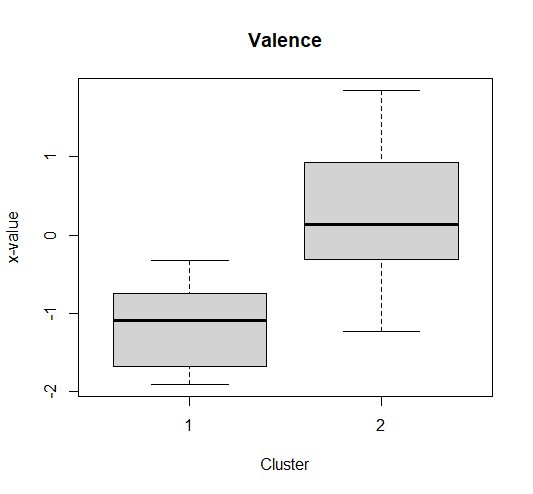
**K-Means Clustering**

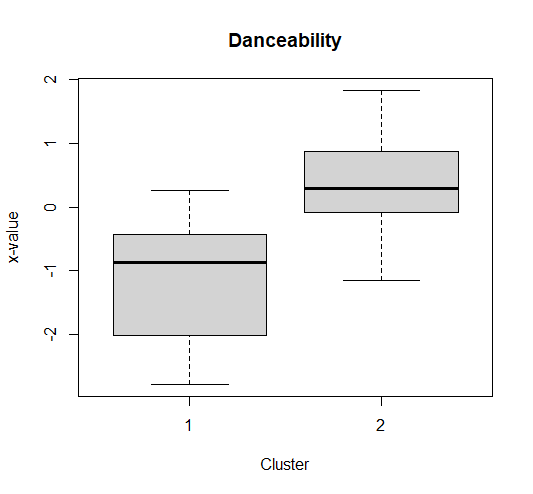
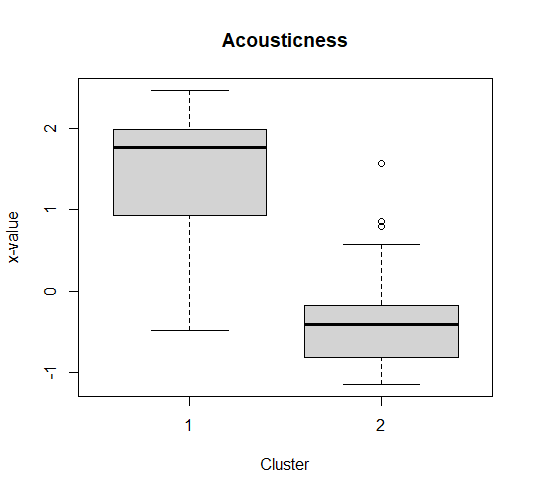
We used K-means Clustering to find common themes within the most popular songs. We focused on songs with a popularity score over 90 for this clustering. The dataset is already clean without any missing information. After scaling and standardizing, we implemented k-means in R with the kmeans function and used the elbow chart to find the optimal number of clusters. The resulting chart is shown below. We determined two clusters to be optimal for our study.





We visualize the 2 clusters by using fviz\_cluster function in R. Since we are focusing on the popularity score over 90, the length of the data contains only 43 songs after filtering in R. From the graph, we can see a majority of the songs belong in the cluster 2 where the cluster 1 has much less songs in the group. This is an interesting visualization to draw down some insights from it. In the blue cluster, most of the songs are upbeat/party songs, but the red cluster is more slow-paced/serious songs. Although both clusters have high popularity scores, most users prefer the musical positiveness of a song being more euphoric, upbeat, and cheerful.





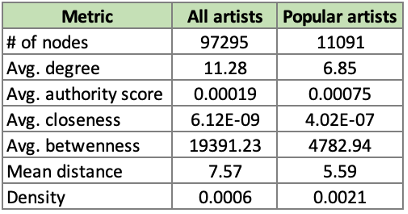
From the boxplots, we conclude that the cluster 2 has more valence, loudness, and danceability where the cluster 1 has higher duration of time and acousticness. Based on these illustrations, all data variables are profiled for each of the chosen clusters. Thus, we understand that if a song gets popular then it is most likely ending up with either upbeat or serious song.

**Network Analysis**

To investigate the driving forces behind what makes a popular song further, we performed network analysis on Spotify’s song dataset. Since people often rely on networks to find information, we thought the artist network may play a role in the exposure, and ultimately the popularity, of a song. We defined our nodes to be artists and a connection to be formed when two or more artists collaborate on a song together. This connection is undirected. Below is an example of a song which was made by two artists in our data. This creates an undirected connection between artists Kendrick Lamar and SZA.

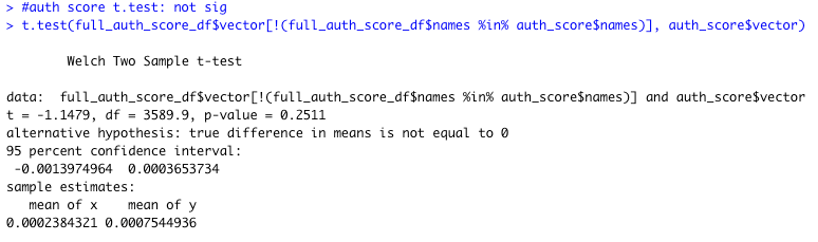


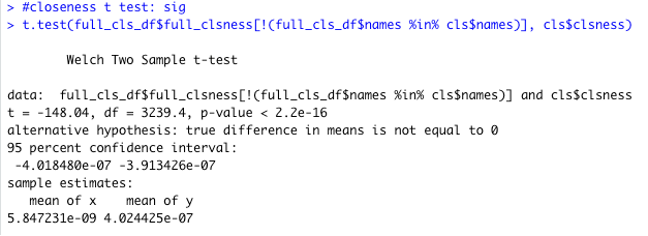
In order to better understand the structure of the artist network, we wanted to compare the entire network to the subset of popular artists. We defined an artist to be popular if they made at least one song with a popularity rating of 70 or more. With our popular artist definition in place, we were able to compare the average degree, authority score, closeness, betweenness, mean distance, and density of the entire dataset as well as for the popular artist network. Below are the results.

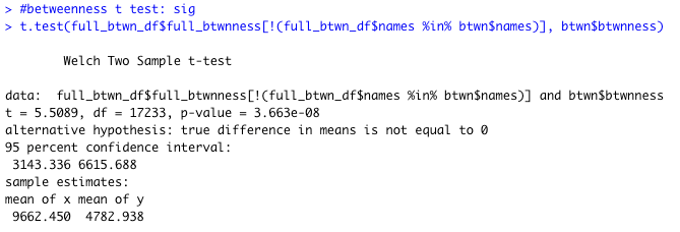


We can see from the table above, the popular artist network has a higher authority score, closeness, and density. Since authority score is the number of incoming degrees (which in our case is the same as the hub score because the graph is undirected), closeness is the inverse of the distance to each node, and density is the number of connections divided by the total number of possible connections, we observe that the popular artist network is more closely knit than the general network, on average. We see this same theme carry over into the lower betweenness and mean distance values as well. The average degree for all artists is higher than the average degree for popular artists. This could be due to the fact that the number of nodes in the popular artist network is about 11% of the full data; hence there are less artists to connect with.

Observing the differences in means is informative, however we wanted to test if the differences in authority scores, closeness and betweenness were statistically significant. To do this, we added a binary flag in our data for popular artists and non-popular artists. We were able to perform t-tests by our popularity flag. The results were as follows,

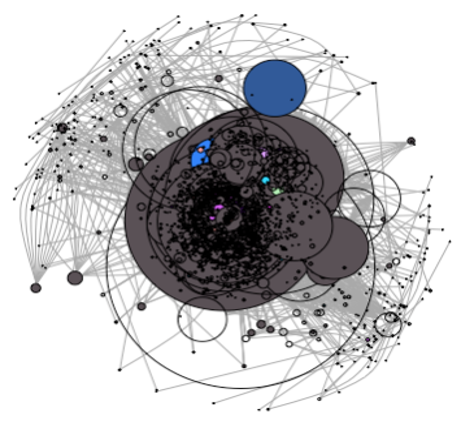
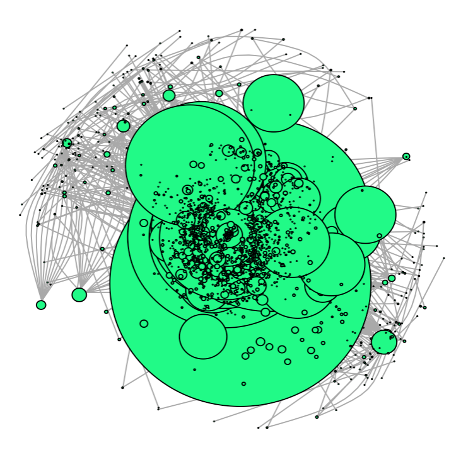






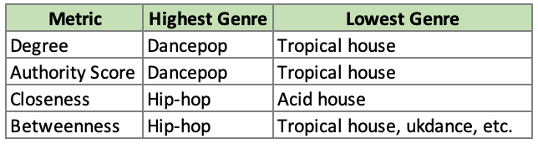
We can observe that the difference in means for closeness and betweenness are all statistically significant, while the difference in authority scores between popular and non-popular artists was not statistically significant. From the higher closeness value in the popular artist community, we can infer that this network is a “smaller circle” than the non-popular community. From the lower betweenness value in the popular artist network, we can infer that popular artists are less likely to broker an exchange between two different artists. This may be explained by the fact that the popular artist network is well-connected and thus an artist is not required to bring other popular musicians together.

The graph below on the left is the visualization of the popular artist network where the size of the node is a function of its degree. The image aligns with the metrics we calculated. The nodes are congregated in the center of this kamada kawai-style graph. We also see a lack of “islands” in the graph. The nodes appear to be connected to one another. It is also evident that there is a high number of connections in this network. These observations follow the same conclusions we developed from the t-tests and average means. The popular artist network is a small, well-connected community.

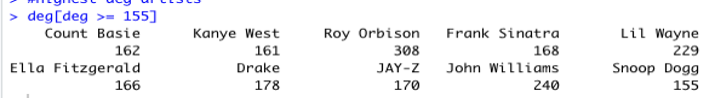


The chart above on the right is a visualization of the popular artist network where the size of the node represents the degree, however the colors represent the artists genre. From this chart, we have visual evidence of homophily. The grey color represents hip-hop artists, and we see a high proportion of the network consists of hip-hop artists. We also see that hip-hop artists are clustered together.

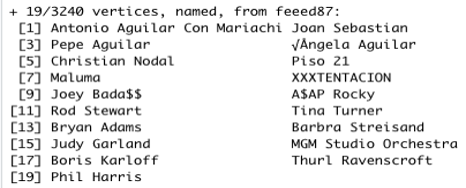
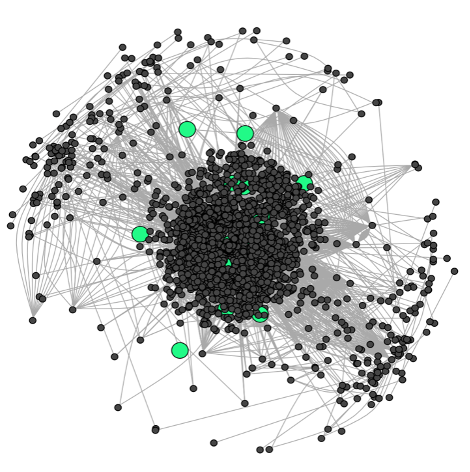
Because we found evidence of homophily in genres, we took a deeper look at the network metrics by genre. Below is a table of the genres with the highest and lowest degree, authority score, closeness, and betweenness values on average. Dancepop (ex. NSYNC) has the highest degree and authority score. This means dancepop is the most common genre for collaborations. Since dancepop is a highly commercialized genre, perhaps some collaborations are a doing of the record labels. Tropical house, however, is the genre with the lowest degree and authority score. This genre (ex. Thomas Jack) is more likely to create songs on their own. For closeness, hip-hop had the highest value. In hip-hop, it is common to sample other artists work and collaborate with various artists, so it makes sense that this genre is the most central. Acid house is the genre with the lowest closeness value. After doing some research, acid house is a subgenre of house music from Chicago in the 80’s. It is a fairly obscure genre which may be the reason for the low closeness score. For betweenness, hip-hop was the genre with the highest value, likely for the same reason as closeness. There were a few genres with the lowest betweenness value of 0. Again we see tropical house ranking the lowest. This may be due to the fact that as a genre, they do not collaborate often and therefore would not connect two artists as much as other genres.



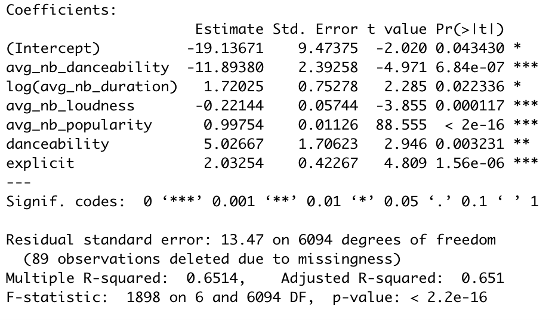
Next, our team took a look at the artists with the highest degree. Below are the top ten artists based on degree. We see that six of the ten are current artists. Perhaps it is easier now for artists to collaborate on music with recording being digital and cheaper to produce. Roy Orbison is the artist with the highest number of connections, beating second place by 68 connections.



We also studied the diameter of the popular artist network. Below is the visualization along with the 19 artists that comprise the diameter. On one end of the diameter, we haveAntonio Aguilar Con Mariachi, a mariachi band from the 50’s. On the other end, we have Phil Harris, an artist from the 30’s most notable for his works in the original Disney Jungle Book cartoon. These two artists are the most polarizing artists that are connected to one another. What our group found particularly interesting in the diameter, was the fact that we start in the 50’s with a mariachi band and within a six more artists, we get to current hip-hop artists. Then we migrate back to artists from the 80’s and end with an artist from the 60’s. After considering the network metrics by genre, however, this makes more sense. As a genre, hip-hop has the highest betweenness score, so it makes sense that hip-hop artists would be found towards the center of the diameter.



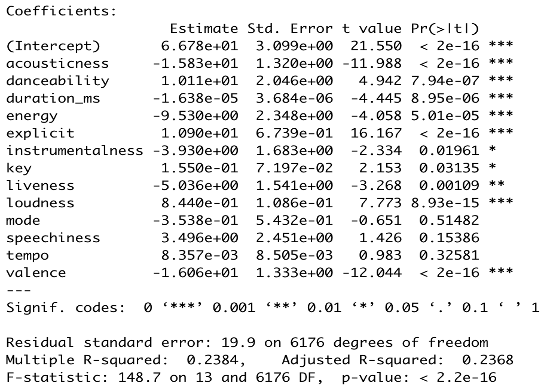
From our network analysis, we see definite trends in the popular artist community. It makes logical sense that the network would be a factor in the popularity of a song. We tested this by creating a linear regression model to predict the popularity of a song based on the songs profile, the main artists’ network metrics (closeness, betweenness, etc.), and their neighbor’s average song profile. After taking the log of skewed variables, removing insignificant variables, and removing variables with multicollinearity, we arrived at the following model,



It is evident that the song’s profile makes an impact on its popularity. An increase of one unit in the danceability score is associated with an increase in the popularity by 5.03, making it more popular. If a song is explicit, it is associated with an increased popularity by 2.03. We can see that the neighbor’s average song profile, and thus the network, is important in the popularity of the song, namely the average neighbor danceability, log of the average neighbor duration, average neighbor loudness, and average neighbor popularity. The interpretations are as follows,

* Average neighbor danceability: An increase in one unit in the average danceability in neighbors is associated with a decrease in popularity by 11.89
* Log of the average neighbor duration: A one percent increase in the average neighbor duration is associated with a 0.017 increase in popularity
* Average neighbor loudness: An increase in one unit in the average loudness in neighbors is associated with a decrease in popularity by 0.22
* Average neighbor popularity: An increase in one unit in the average popularity in neighbors is associated with an increase in popularity by 1.00

To further the point that the network is an important factor in a song’s popularity, we ran another linear regression model based solely on the song’s profile. The results are below. The model below based on only the song’s profile has an r-squared of 0.237 compared to the 0.651 when we consider network influences.



**Conclusion**

There are a number of business implications from our network study. On the artist and record labels’ sides, they should leverage their peers’ audience to ensure a song is popular and emphasize the two main segments of popular songs, upbeat party songs and slow-paced serious songs. Spotify could also leverage this in their recommendation systems. If a song is released, and it fits our ideal profile and placement in the network, Spotify can preemptively add it to its automated playlists or place it towards the top of the “New Music Friday” playlist. Another insight we found interesting was that danceable songs tend to be more popular than non-danceable songs. However, a neighbor's danceability has a negative effect on the song’s popularity, meaning the more danceable the neighbor is, the less popular the song is. This could be the impact of dance remixes to already popular songs. These tend not to be nearly as popular as the original song but are commonly placed on the user-specific automated playlists. This should be a point for Spotify to look into and update their algorithm if it is found to be true.