Project Report

MATH 448

Nguyen Ly

5/3/2021

Introduction

Research Topic

As streaming is becoming the main way of interacting with music, this creates changes to how music is consumed. This project is an analysis of music recommendation and interaction, specifically on Spotify. Music recommendation and interaction will be examined using public playlists made by Spotify users. This project will further delve into the success of a track and its popularity on the streaming platform based on playlist recommendation and interaction. The metrics for measuring music recommendation and interaction is through the success of artist and track with popularity, playlist count, and follower count. These are variables in which music recommendation and interaction would increase the value. This project will utilize data from the Spotify Million Playlist Dataset Challenge in which Spotify released datasets on 1 million playlists. However, while there is some randomization, this dataset is not representative of the distribution of all datasets on the Spotify platform due to filtering for quality and removing offensive content. The second source of data comes from Kaggle which will integrate more variables into the Spotify dataset. The data from Kaggle is a collection of around 600k tracks from 1922 to 2021 which a more in-depth look at a track's variable such as danceability, energy, etc.

Data Description

The first set of data is from Kaggle (see source below). It is a collection of around 600,000 tracks collected from Spotify Web API and was last updated on 4/20/2021. The two datasets taken from this source are artist.csv and tracks.csv. Both files include variables that are categorical, numerical, and ordinal. The first file, artist.csv includes has 5 variables. There are 3 categorical variables which include id, name, and genre, and two numerical variables which are followers, and popularity. The second file, tracks.csv, has 7 categorical variables which are id, key, timesignature (overall time signature of a track, with how many beats are in each bar), artists, artists(id of the artist), release date, and name(the name is the song). There are 11 numerical variables which are acousiticness (confidence measure from 0 to 1, with 1 being high confidence that a track is acoustic), danceability (measure from 0 to 1 on how suitable a track is for dancing based on its tempo, rhythm stability, beat strength, and overall regularity), energy (measure from 0 to 1 on perceptual measure of intensity and activity), duration_ms (duration of a song in ms), instrumentalness (confidence measure from 0 to 1 as to whether a track is instrumental or vocal, with 1 representing intrumental tracks), valence (measure from 0 to 1 describing musical positiveness, with 1 being more happy and cheerful tracks), popularity (popularity score of a track ranging from 0 to 100), tempo (overall temppo of a track in beats per minute), liveness (detection of audience in recording, with 0.8 being a strong likelihood a track is live), loudness (overall loudness of a track in decibels), and speechiness (measure from 0 to 1 representing spoken words in a track, values between 0.33 and 0.66 describing tracks containing both music and speech). There are two ordinal variables which are mode (modality of a track, 0 being minor and 1 being major) and explicit (whether a track contains explicit content, with 1 being yes and 0 being no). Each song is distinguished by a unique ID generated by Spotify. Explanation of the variables can be found at this link: https://developer.spotify.com/documentation/web-api/reference/#endpoint-get-audio-features.Changes were made to the original dataset to change duration_ms to duration_min (which shows the duration of a track in minutes).

The second set of data is from the Spotify Million Playlist Dataset Challenge. The dataset was first released in 2018 which includes a small sample of 1 million playlists from a population of 4 billion public playlists on Spotify. These playlists were created by Spotify users between January 2010 to November 2017. This dataset is not a great representation of the distribution of playlists available on Spotify since it was "manually filtered for playlist quality and to remove offensive content and have some dithering and fictitious tracks added to them". In total, there are 18 variables for this dataset, with 10 variables about the playlist and 8 variables about the individual tracks. Categorical playlist variables include name, and collaborative. For the playlist, there are 10 variables, with 7 of the variables being numerical and 3 of the variables being categorical. The numerical variables are pid (ID number assigned to each playlist in the dataset), num_albums (numbers of unique albums in the dataset), num_tracks (number of unique tracks in the dataset), num_followers (numbers of followers a playlist has), num_edits(number of edits the user has done on the playlist), duration_ms (how long the playlist is in ms), and num_artists (numbers of unique artists in the playlist). The categorical variables are name, collaborative (is the playlist collaborative, with 1 meaning yes and 0 meaning no), and modified_at (last modified date for the playlist). For each track on the playlist, there are 8 variables with 2 numerical variables and 6 categorical variables. The numerical variables are pos (position of the track in the playlist) and duration_ms (how long the track is in ms). The categorical variables are artist_name, track_uri (unique identification string of a track), artist_uri (unique identification string of an artist), track_name, album_uri, and album_name.

Both datasets were merged using artist uri or track uri which were found in both datasets. After removing rows with empty cells and missing data, the dataset for tracks went from around 2.2 million unique tracks to around 100,000 tracks. For the dataset for artists, there were cases where the artists were repeated due to having collaborations or having labeled with multiple artist uri. After cleaning, the data went from having 298,000 artists to around 113,000 unique artists.

Research Questions:

1. What variables increase the popularity of a song?
2. What is the effect of a track's popularity on the number of Spotify playlist a track is in?
3. Does having a high playlist and follower count influence an artist's popularity, assuming that popularity score is an indicator of interaction with an artist?

Data

- https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks
- https://www.aicrowd.com/challenges/spotify-million-playlist-dataset-challenge#dataset

Statitical Results

What variables increase the popularity of a song?

A song's popularity on the Spotify platform describes its success in streaming. The song's popularity score on Spotify is valued from 0 to 100, with 0 being the least and 100 being the most popular. This value is generated by an algorithm based on the number of plays the track has and how recent are the streams.

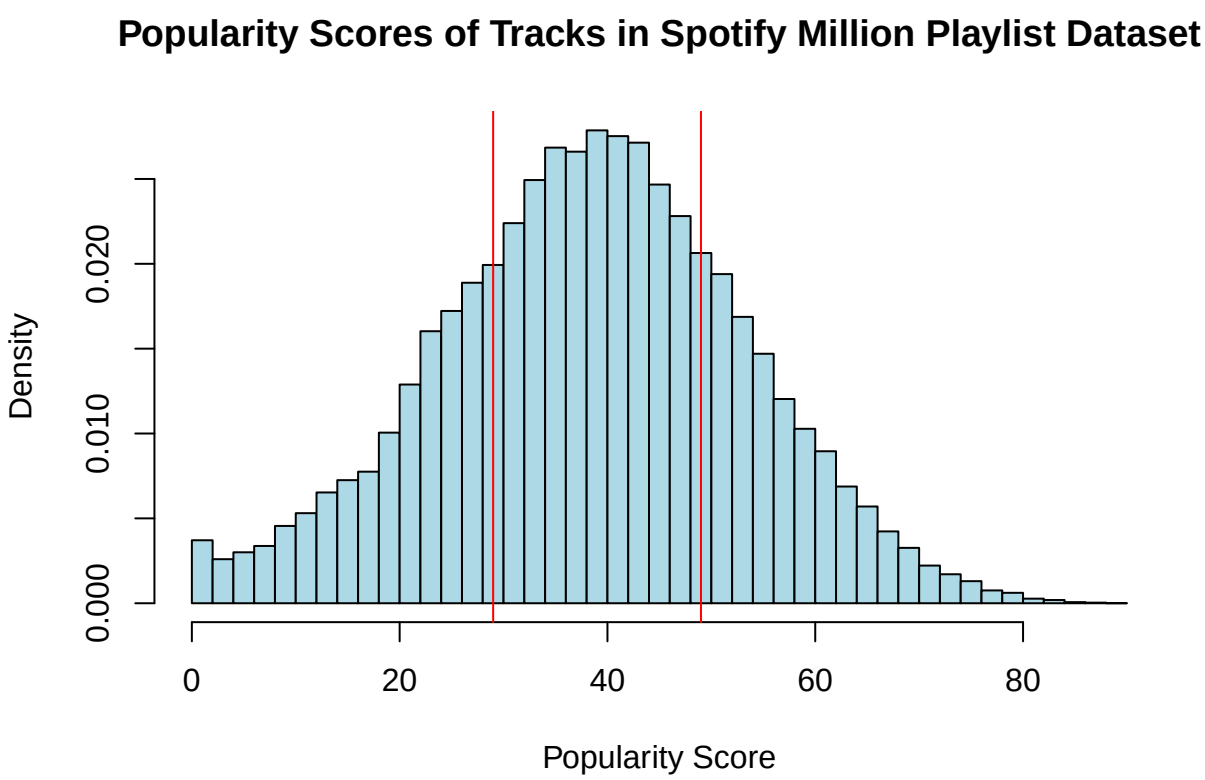

Figure 1: Histogram showing distribution of popularity score of tracks in the Spotify Million Playlist Dataset. The two red lines indicate quartile 1 and quartile 3 respectively.

After merging the data from Spotify Million Playlist Dataset and Kaggle, and omitting data points with missing values, Figure 1 was produced showing the distribution of popularity scores of tracks in Spotify Million Playlist Dataset. The distribution of the histogram is normal and unimodal, with a mode at 40 points. It has a minimum value of 0 points and a maximum value of 90 points. The first quartile is at 29 points. The third quartile is at 49 points. The median and mean are 39 and 38.92986 respectively.

To understand what variables influence the popularity of a track, we have to look into the correlation of each variable with respect to the popularity of a track. Figure 2 shows multiples plots of popularity versus the 11 numerical variables such as acousticness, count (number of playlists a song is in), danceability, duration, energy, instrumentalness, liveliness, loudness, speechiness, tempo, and valence.

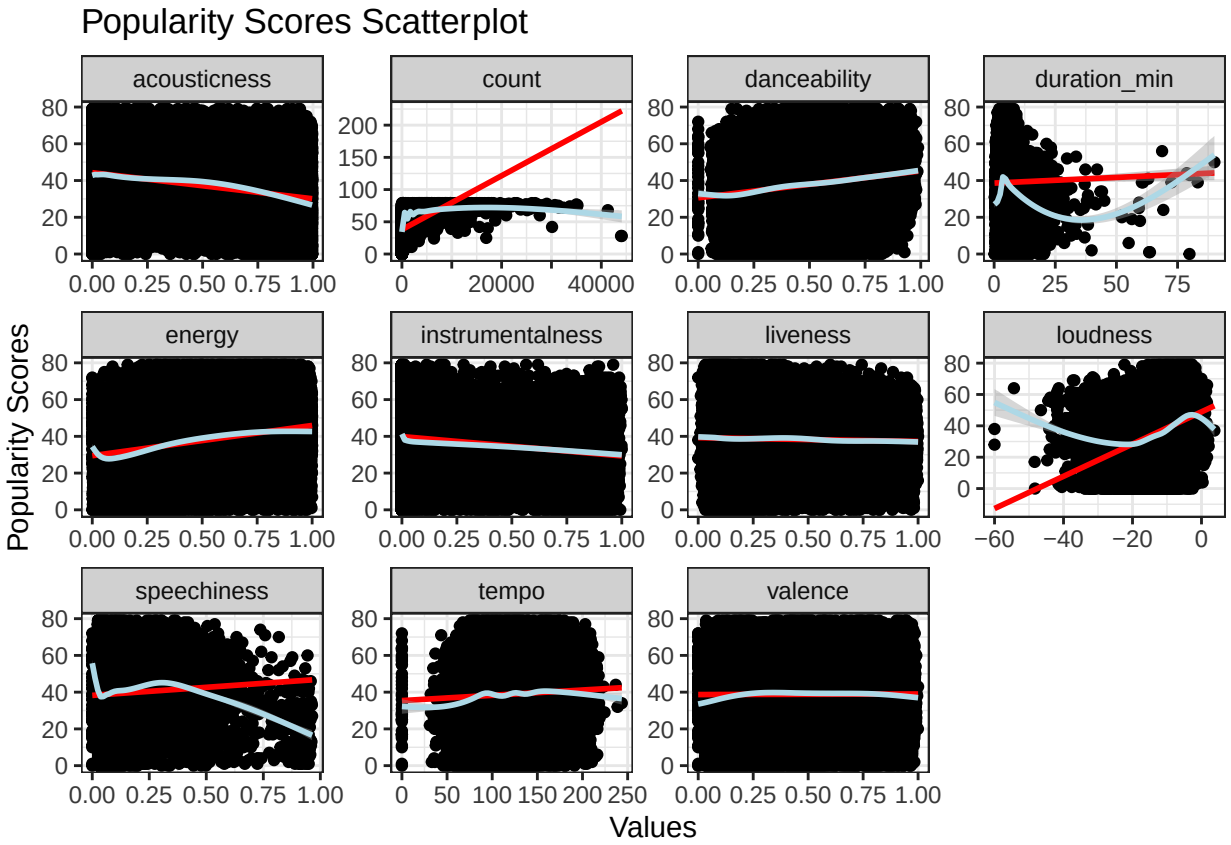

Figure 2: Scatterplots of popularity scores vs 11 numerical variables. The blue line represents the generalized additive model and the red line represents the linear model.

From the figure, there seem to be certain patterns in how each variable influences popularity. There are 6 variables that could be modeled as linear relationships such as acousticness, danceability, instrumentalness, liveness, tempo, and valence. Since a 1.00 represents higher confidence that the track is acoustic, as acousticness increases to 1.0, there is a decrease in popularity to around 25 points. With danceability, there is an increase in danceability to 1.0 as popularity increases from 30 points to around 50 points. For instrumentalness, there is a decrease in popularity as instrumentalness increases. When instrumentalness was at 0, popularity is around 40 points. As instrumentalness increases to 1, popularity decrease to around 30 points. For liveness, there is little to no change to popularity as liveness increases. For tempo, there is a slight increase in popularity as tempo increases from 0 to around 100 BPM. However, popularity does not increase when the tempo increases from 100 to 250 BPM. For valence, there is a slight increase in popularity by 10 points when valence increase from 0 to 0.25. As valence increases from 0.25 to 0.8, there was no change in popularity. However, when valence is between 0.8 and 1.0, popularity decreases slightly.

For count, duration_ms, energy, loudness, and speechiness, a linear relationship is not well fitted to the data. The relationship between follower count and popularity score will be examined more in-depth in the next section. However, there tends to be an increase in popularity score as the follower count increases. As for the duration of a song in comparison to popularity, there is an increase in popularity as a track's duration increases from 0 to 4 minutes. As for duration increases from 4

minutes to 35 minutes, there is a decrease in popularity by 20 points. As for duration in minutes increases from 35 minutes towards 91 minutes, popularity increases by 30 points. As for loudness, there is a decrease in popularity by 30 points as loudness increase from -60 to -20 decibels. There is an increase in popularity by 20 points as loudness increase from -20 to -8 decibels. From -8 to 0 decibels, popularity decrease by 10 points. For speechiness, as speechiness increases, popularity decreases. As speechiness increases from 0 to 0.04, popularity decreases by 20 points. There is a slight increase in popularity by 10 points as speechiness increases from 0.04 to 0.33. As speechiness increase from 0.33 to 1.0, popularity score decrease by approximately 30 points.

From the figure, there is a positive linear relationship between popularity and 6 numerical variables. These are count, danceability, duration_min, energy, loudness, and speechiness. While some of these variables are not well modeled to a linear model, this positive linear relationship suggests that as these variables increases, popularity also increases.

Table 1: Correlation between popularity and the 11 numerical variables.

Variables	Correlation to Popularity	Cor.Test P-value	95% Confidence Interval
count	0.3305338	2.2e-16	(0.3322, 0.3425)
popularity	1.0000000	n/a	n/a
danceability	0.1676688	2.2e-16	(0.1626, 0.1739)
energy	0.2778534	2.2e-16	(0.2723, 0.2831)
loudness	0.3451695	2.2e-16	(0.3408, 0.3510)
speechiness	0.0478933	2.2e-16	(0.0421, 0.0537)
acousticness	-0.3172420	2.2e-16	(-0.3228,-0.3123)
instrumentalness	-0.1760148	2.2e-16	(-0.1821, -0.1708)
liveness	-0.0264352	2.2e-16	(-0.0329, -0.0212
valence	0.0054542	0.08202	(-0.0007, 0.0110)
tempo	0.0579349	2.2e-16	(0.0514, 0.0630)
duration_min	0.0077919	0.01657	(0.0013, 0.0130)

To look into the linear relationship between all 11 numerical variables in comparison to popularity, we can look into the correlation between popularity and the other 11 numerical variables in Table 1. There is a moderate degree of correlation between popularity and three other variables: count, loudness, and acousticness. There is a moderate positive degree of correlation between popularity and 2 variables. These are count and loudness with a correlation of 0.331 and 0.345 respectively. Also notable is the correlation between popularity and energy. Although low, it has a correlation coefficient of 0.278. A correlation test was conducted to test the correlation between popularity and the 11 variables. The p-values from the test can be seen in Table 1. For all the variables with the exception of valence and duration of tracks in minutes, the null hypothesis can be rejected at the 1% significance level. This suggests that the true correlation between popularity scores and all the variables (with the exception of valence and duration in minutes) is not equal to 0. For the correlation between popularity scores and valence and duration in minutes, the null hypothesis is accepted at the 1% significance level. The data suggests that for these two variables, the true correlation between popularity scores and these two variables is equal to 0. A 99% confidence interval for all the variables is found in Table 1. It is notatble count and loudness have the highest positive correlation to popularity. However, in the scatterplot of popularity scores, count and

loudness does not follow the linear model well.

What is the effect of a track's popularity on the number of Spotify playlist a track is in?

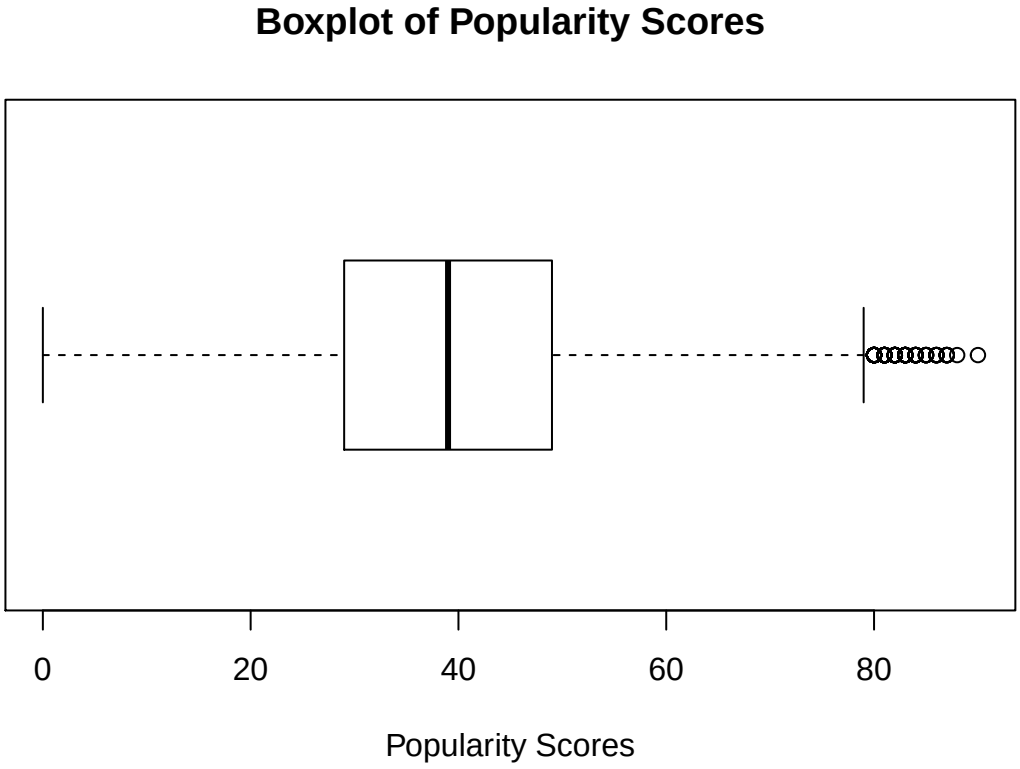

Figure 3: Distribution of popularity scores in dataset

When looking at Figure 2 of a track's popularity and the number of Spotify playlist a track is in, all the outliers consist of tracks that have a higher popularity score (79 points and above). These are tracks that did exceptionally well in comparison to the rest of the dataset. To distinguish how the number of playlists a track is in effect a track's popularity, both variables will be converted into categorical variables with varying categories based on their IQR. Values below Q1 will be categorized as low. Values between Q3 and Q1 will be categorized as mid, and values above Q3 will be categorized as a high or high outlier. For popularity scores, a low score is between 0 and 28 points. A medium score is between 29 to 49 points, and a high score is between 50 to 79 points. The outliers will be kept and will have their level which is between 80 and 100 points. As for playlist counts, a low playlist count is between 1 and 2 playlists. A medium playlist count is between 3 to 37 playlists, and a high playlist count is between 37 to 46574 playlists. Table 2 is a contingency table that displays the frequency of distribution of popularity score and playlist count using the categories discussed above.

Table 2: Contingency table of popularity scores and playlist count, with columns being Spotify popularity scores and rows being playlist counts.

	Low	Medium	High	High Outlier
Low	15917	16910	1959	0
Medium	10599	32331	6824	2
High	426	9883	18017	183

Table 3: Percentage Contingency table of popularity scores and playlist count, with columns being Spotify popularity scores and rows being playlist counts.

	Low	Medium	High	High Outlier
Low	59.08	28.60	7.31	0.00
Medium	39.34	54.68	25.46	1.08
High	1.58	16.72	67.23	98.92

Looking at the percentage contingency table in Table 3, it is notable that 98.92% of tracks with a high outlier score (with a score of 80 or more) will also be put into a high number of playlists. The other 1.08% of tracks with high outlier scores will be put into the mid number of playlists, with 0% being put in a low number of playlists. Similarly, for high popularity scores, 92.69% of songs with high popularity scores being put in the mid to high numbers of playlists. On the other side of the spectrum, 59.08% of low popularity scores are put into a low number of playlists. This seems to indicate some type of dependency relationship between the two variables. Also from the contingency table, we can compute an odds ratio to further examine the relationship between popularity score and playlist count. A track with high popularity scores (which are scores between 50 to 100) is 15 times more likely to have a high playlist count than a track with a mid to low popularity score.

$$\text{Odd Ratio} = \frac{\text{odds of high popularity with high playlist count}}{\text{odds of low popularity with high playlist count}}$$

$$= \frac{\frac{\text{high/high outlier popularity with high playlist count}}{\text{high/high outlier popularity with low playlist count}}}{\frac{\text{low/medium popularity with high playlist count}}{\text{low/medium popularity with low playlsit count}}} = 15.22425$$

From the contingency table, there seems to be some sort of dependent relationship between the two variables. We can utilize the chi-square test of independence to test whether these two categorical variables are related or not. A Pearson's Chi-Square Test was conducted. The null hypothesis is that there is no association between playlist count and popularity score. The alternative hypothesis is that there is an association between playlist count and popularity score. The resulting chi-square test has a $\chi^2_6 = 272.11$ with 6 degrees of freedom. The resulting p-value is $p < 2.2e-16$, which is less than $\alpha = 0.05$. since the p-value is less than alpha, we can reject the null hypothesis assumption of independence, meaning that the popularity score is not independent of playlist count.

Does having a high playlist and follower count influence an artist's popularity, assuming that popularity score is an indicator of interaction with an artist?

Metrics such as playlist count or follower count can indicate a user's interaction with an artist. This could be through following an artist to keep updated with their music or adding an artist to their playlist. Therefore having a higher playlist count or follower count would increase an artist's popularity as well, since this means that more users are interacting and recommending the artist's music. To see how these variables interact and the relationship between these variables, we first look at some descriptive statistics.

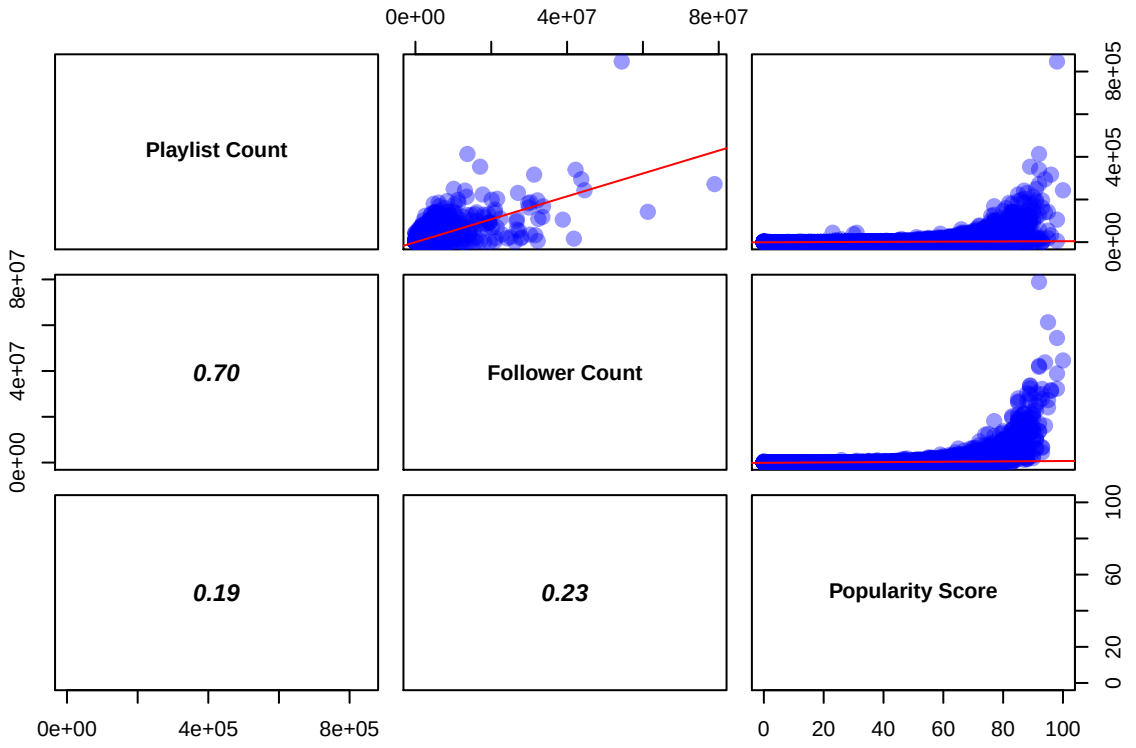

Figure 4: Figure of scatterplot with linear regression and correlation of playlist count, followers count, and popularity score.

From Figure 3, we can see the histograms of the three variables plotted against each other with a linear regression fitted to the plot. While there is a strong linear relationship between playlist count and follower count, the graph of playlist count and the graph of popularity score and follower count and popularity score does not have a strong linear relationship. This relationship can be seen with the correlation coefficient between playlist count and follower count of 0.70, which is a strong correlation. However, the correlation coefficient for playlist count and popularity score and the correlation for follower count and popularity score is 0.19 and 0.23 respectively, which is a low to moderate correlation. For all these correlations between the variables, a correlation test was conducted in which the p-value for all the tests was statistically significant at a significance

level of less than 0.001. This means that we can reject the null hypothesis that there is not a significant linear relationship between these variables. We accept the alternative hypothesis that the correlation coefficients are significantly different from 0.

Table 4: Proportion table of an artist's popularity score versus their playlist count. The rows represent playlist count, and the column represent popularity score.

	Low	High
Low	0.8678	0.3853
High	0.1322	0.6147

Table 5: Proportion table of an artist's popularity score versus their follower count. The rows represent follower count, and the columns represent popularity score.

	Low	High
Low	0.921	0.2149
High	0.079	0.7851

Knowing that there is a correlation between these three variables, we can build a contingency table of these variables to look at them in terms of categories since the data is so immense. The categorization of the variables in this section is different from the previous. A low popularity score, follower count, and playlist count are values below Q3. A high popularity score, follower count, and playlist count are values above Q3, which also include outliers in this dataset.

To measure success as an artist based on popularity score, we can look at the relationship of having a low or high popularity score versus playlist count and follower count. From the contingency table, we can see that artists with high popularity are more likely to have a higher follower count and playlist count, with a probability of 0.7851 and 0.6147 respectively. Artists with a high popularity score are 10.47 times more likely to have a high playlist count than artists with low popularity scores. This means that artists with high popularity scores are included in a higher number of playlists compared to artists with low popularity scores. The relative likelihood for an artist with a high popularity score to have a high playlist count is 9.94 times the likelihood of an artist with a low popularity score to have a high playlist count. Artists with a higher popularity score are 42.59 times more likely to have a high follower count than artists with a low popularity count. The relative likelihood for an artist with a high popularity score to have a high follower count is 4.65 times the likelihood that an artist with a low popularity score to have a high follower count.

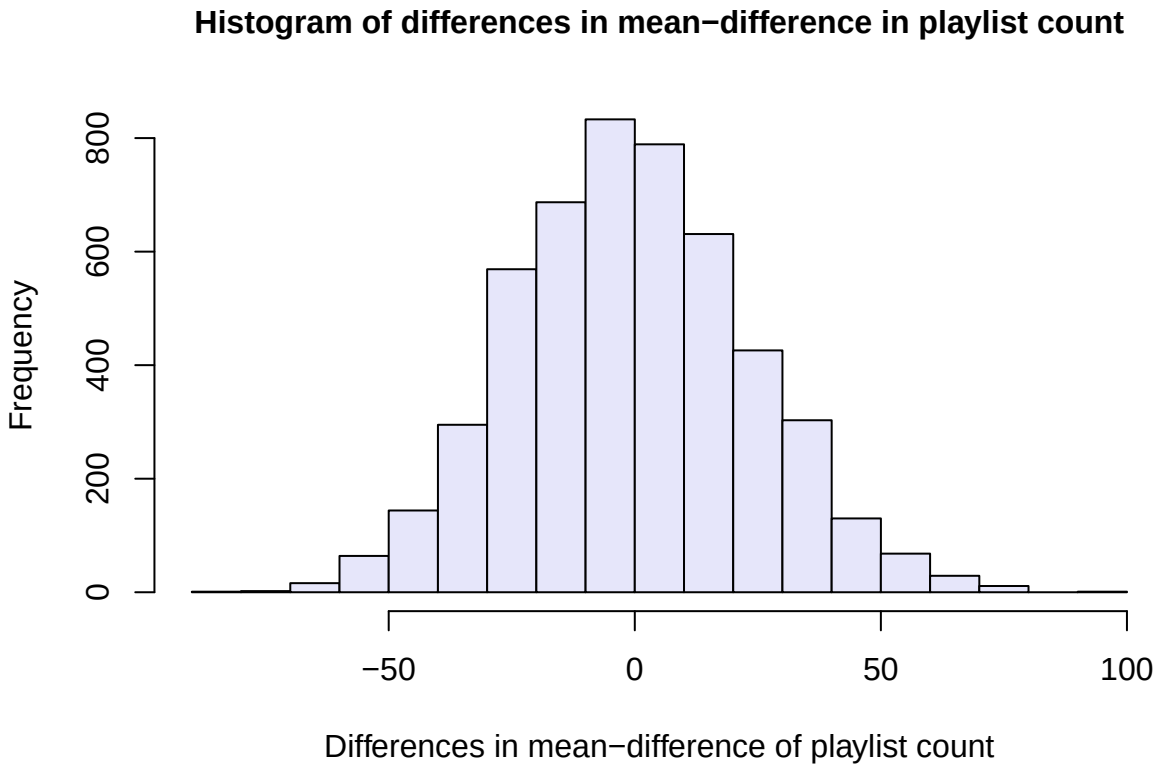

Figure 5: Histogram of differences in mean-difference in playlist count and popularity score of an arist.

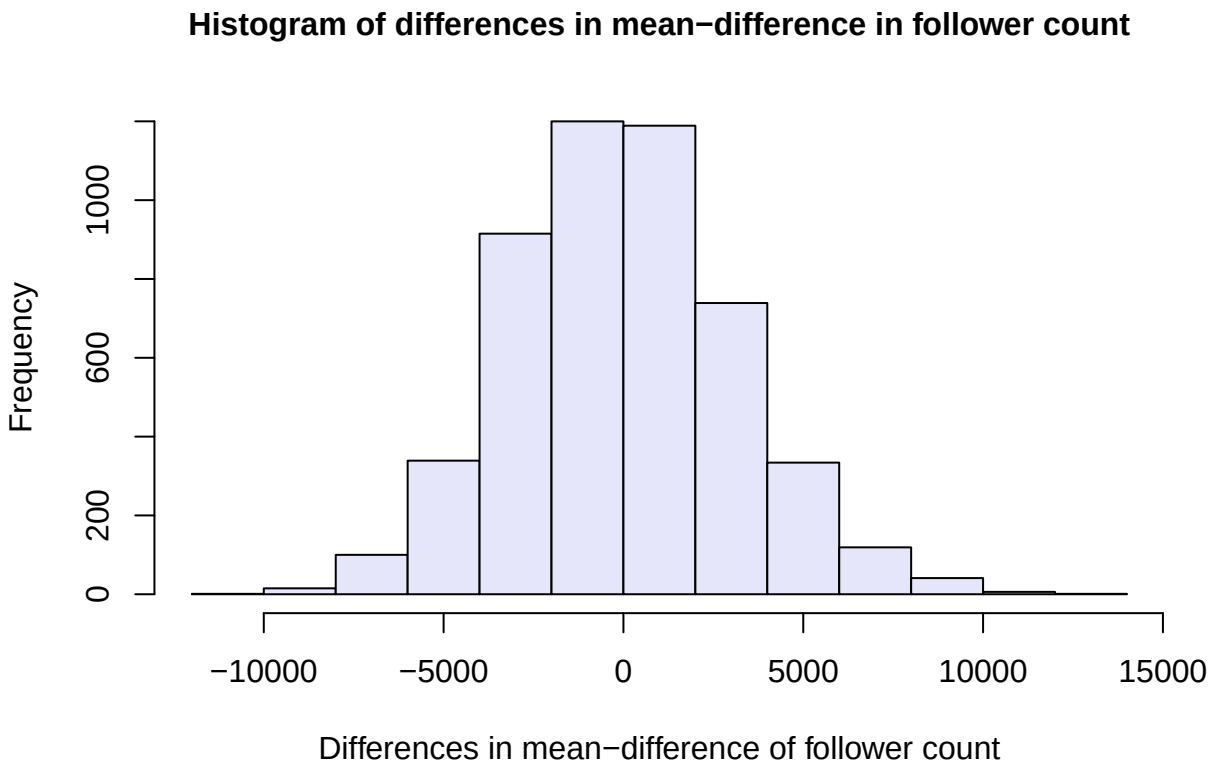

Figure 6: Histogram of differences in mean-difference in follower count and popularity score of an arist.

While the relative likelihood and odds ratio does describe the relationship between follower count and playlist count to popularity scores, we have to test if the differences in these proportions are equal. To do so, we generate a permutation test. The null hypothesis for the first test is that the probability of having a high playlist count is the same between artists with low popularity scores and artists with high popularity scores. The alternative hypothesis is that the probability of having a high playlist count is higher for artists with high popularity scores than an artist with low popularity scores. The differences in mean-difference of playlist count can be seen in Figure 4. While lines were added to show the difference in means of the dataset (which were at 1272.685 and 1272.685)onto the resampling from the permutation test, the lines were outside the distribution of the permutation test. The p-value for this test is $2e-04$ meaning that the p-value is statistically significant. We can reject the null hypothesis in support of the alternative hypothesis that the probability of having a higher playlist count is higher for artists with high popularity scores than artists with low popularity scores.

The second permutation test is conducted to compare proportions in follower count and popularity scores. The null hypothesis is that the probability of having a high follower count is equal for artists with high popularity scores and artists with low popularity scores. The differences in mean-difference of follower count can be seen in Figure 5, where the resampling of differences in mean-difference have a distribution between -10,000 and 15,000. However, the difference in mean-

difference for the dataset is at -204527.6 and 204527.6 which is not included within the resampling range. The p-value for this permutation test is at $2e-04$ meaning that the p-value is statistically significant. We reject the null hypothesis in support of the alternative hypothesis that the probability of having a higher follower count is higher for artists with high popularity scores than artists with low popularity scores.

Conclusion

Summary of Findings

- Of the 11 numerical variables examined in comparison to popularity, there is a positive linear relationship between popularity score and six variables: playlist count, danceability, duration in minutes, energy, loudness, and speechiness. Of those six variables, playlist count and loudness have the highest value of positive correlation to popularity, with a correlation coefficient of 0.331 and 0.345 respectively.
- A track that has a high popularity score ranging from 50 to 100, is 15 times more likely to have a high playlist count than a track with a mid to low popularity score.
- Based on the Chi-Square Test of Independence, the p-value is statistically significant. This meant that the p-value was less than the alpha level. Therefore, the popularity score is not independent of playlist count.
- Having a high popularity score also mean that an artist will have a higher probability of having a high follower count and high playlist count, assuming follower count and playlist count are metrics for user's interactions with an artist.
- While there is a high correlation between follower count and playlist count, there is a low correlation between follower count and popularity score, and playlist count and popularity score. Although these correlations are low, the correlation test reveals that the correlation between these variables is statistically significant. This meant that the correlation between these variables was not 0 and there is a relationship between these variables.
- Artists with a high popularity score are 10.47 times more likely to have a high playlist count than artists with low popularity scores. The relative likelihood for an artist with a high popularity score to have a high playlist count is 9.94 times the likelihood of an artist with a low popularity score to have a high playlist count. Artists with a higher popularity score are 42.59 times more likely to have a high follower count than artists with a low popularity count. The relative likelihood for an artist with a high popularity score to have a high follower count is 4.65 times the likelihood that an artist with a low popularity score to have a high follower count.
- Through the permutation test, we determined that the probability of having a higher playlist count is higher for artists with high popularity scores than artists with low popularity scores, and the probability of having a higher follower count is higher for artists with high popularity scores than artists with low popularity scores.

Future Work

Now that I can utilize certain variables in the data to use as metrics to gauge user interactions and recommendations, I can use this information to delve further into questions such as:

- What does the average position of a track in a playlist mean for user interaction? Does tracks in lower position (first in playlist) in the playlist have higher popularity scores (meaning that people are more likely to interact with it)?
- What does this information mean for the demographic of users who made these playlists? (74
- What does artists with higher popularity scores have in common, considering variables related to tracks such as energy, valence, etc.?
- Is there a bias towards certain genres on Spotify? Especially since most of the popular artists are within genres such as pop and hip-hop? (Consider additional dataset such as BillBoard rankings)