Evaluating Song Popularity Over Time Project Report

Lucas Henry, Philip Nguyen, Tanner Bye, Jacob McKelvy

BANA 471 Data Exploration & Visualization

3/15/2024

Oregon State University

Table of Contents

Introduction

What components of a song make it popular? By evaluating different components of a song, you can assist in promoting a song's popularity in order to provide artists with greater hits. The issue of promoting a song's popularity is multipolar and is a relevant one to address because of a number of reasons. User engagement is important to music streaming platforms and learning what makes a song popular can help a company create playlists that are closer to what users are looking for. By providing music that would be more similar to a customer's taste, you can retain more customers. Additionally, it can make more accurate personal recommendations as music streaming platforms would then know what the interests of listeners are. Utilizing customer interests, you can keep track of current trends in order to further assist artists in promoting their music. Furthermore, it can assist in promoting various newer artists who upload their music to the platform expanding on the interest of artists wanting to utilize the platform. What this all amounts to is that it helps grow the customer base of the streaming platform, which in turn would lead to increased revenues. By utilizing a dataset of the past century's worth of songs, we can give statistically proven evidence on what components of a song make it popular.

Data Cleaning

To deal with duplicates, we first dropped all the duplicates, then dropped duplicates based upon name and artist. The reason for dropping based on name and artist was because an artist could have released the exact same song multiple times. Take Gina Star's song 1000 Years - BARE Remix, for example. They released this song 5 times in 2013 and in each occurrence, the song had the exact same data for all columns except for ID. Since these were the same songs, we just decided to drop all of the duplicates so we weren't measuring the same song multiple times. Initially, the dataset started with 174389 values across all columns. After the first iteration of dropping duplicates, we were left with 172230. Then, after dropping based on name and artist, we were left with 159441.

Additionally, we needed to standardize string values to lowercase. All of the song and artist names are a mixture of uppercase and lowercase letters which makes it difficult to analyze. We converted these to lowercase to help standardize the data.

There are zero null values across all columns, however, there are 121 songs that have 0 tempo so we needed to find a way to deal with those null values. If we were to drop those values, then the minimum would be 51.316. This signals that these songs are significant outliers. Logically, it also doesn't make sense for a song to not have a tempo. Perhaps that means that there are some values missing that had 0 inputted in its place. Either way, we decided to drop these songs to not skew the data.

We also dropped all rows with speechiness values of 0.66 and above. This is because, based on the definition of speechiness provided, anything with above a 0.66 speechiness value was highly likely to be a podcast, audiobook, or talkshow. Since we are analyzing songs, we decided to exclude these rows from our dataset.

Lastly, we decided to drop all values in the year 1920 as there were many songs released in later years that were recorded in the dataset as 1920. An example of this would be Jodoli's son 96.96.96 which although listed as released in 1920, has all of their other albums released in 2018 or later. There are many more examples of this and instead of going through each song and artist individually, we decided to just drop this year. After checking a small sample of songs in other years, this didn't seem to be a recurring problem.

Data Analysis

In order to determine what components of a song would determine popularity, we decided to run a linear regression. The reason we ran a linear regression analysis for the business problem is because it allows for the assessment of testing different variables and how much it correlates and affects popularity. To elaborate, regression analyses are useful because they allow the chance to experiment with different variables and understand which variables are more significant in affecting/predicting the dependent variable. Remember that in the business problem, we wanted to learn what variables would help make a song popular. First, the regression analysis requires that we select a dependent variable to test - which in our case was popularity. In order to keep track of trends and break up the century's worth of data in our dataset, we decided to segment the data into four separate quarters to provide insights for how popularity has changed in songs over time. These quarters are in segments of 25 years: quarter one 1921-1946, quarter two 1947-1971, quarter three 1972-1996, and quarter four 1997-2021.

After completing the initial regression analysis, we found five total independent variables to focus on that were statistically significant across all quarters. We needed this requirement because it creates difficulty in analyzing trends over time if the significant variables constantly change. The five independent variables were danceability, speechiness, acousticness, liveness, and instrumentallness. From here, we further narrowed it down to the two most positive and negative influences based on the coefficients overall leading us to exclude liveness from our data. For our positive variables, we determined that danceability and acousticness have the most positive influence on a song's popularity. Including higher values of these variables can benefit a song's popularity. On the other hand, we found that Speechiness and instrumentallness have the most negative influences on a song. Higher values of these can lead to a song becoming less popular. Looking at each variable more closely we explain why this is and how the variables affect the songs.

Lastly, while the R^2 is an important measure of correlation between the independent and dependent variables, we found that its usefulness was limited due to the high variability in song popularity over the years and the complexity of having many attributes that define a song. There isn't a single variable that can determine popularity by itself as a song requires many aspects of definition. Our R^2 value in our tests never reached higher than 25.6 percent which suggests that while our variables may be individually significant, they collectively don't do a great job at providing a strong predictive power for song popularity.

Instead, we prioritized examining the coefficients from our regression analysis. These allow us to see insights on the general trends in the interests of the listeners. While the R^2 value quantifies the overall fit of the model, the coefficients enable us to look at each variable individually and identify the attributes that are preferred among listeners. This approach provides a more nuanced understanding of the factors that drive popularity.

Popularity

	Q1	Q2	Q3	Q4	Overall
Std Dev	5.608	15.3616	11.319	27.1358	21.815
Min	0.000	0.000	0.000	0.000	0.000
Max	80.000	87.000	90.000	100.000	100.000
Average	1.898	18.344	38.000	35.605	26.971

Popularity (dependent variable) is the main variable that we are referencing when we are trying to predict what makes a song popular. It is a current measure of popularity so if a song was released in 1920, the popularity isn't a measure of popularity in that year - it's the popularity of that song in the present day. This is important when referring to the numbers we have listed. As we can see, the average popularity is far less in the first two quarters than the last two with the max popularity increasing as time goes on (see Figure 1).

Q1	White Christmas, Bing Crosby
Q2	Rockin' Around the Christmas Tree, Brenda Lee
Q3	All I Want for Christmas is You, Mariah Carey
Q4	Drivers License, Olivia Rodrigo

As noted in the table above, we can see that when we break up our analysis into four quarters, the first three quarters are all similar in the fact that they're just Christmas music. This is indicative of our dataset being skewed based on today's interests and not the interests of the time. Since most people nowadays aren't listening to older songs unless it's for a specific reason like the holidays. Quarter four however, has the song Driver's License as the top song. Driver's License is a pop song that was released by Olivia Rodrigo in 2021, the year of this dataset was released. This is a song that we can confidently determine was impacted based on user interests rather than just being related to a random time of year. From this we were able to decide to focus the majority of our research on this quarter as it would have the most accurate results as well as explain more about current trends.

Danceability

Danceability is a key independent variable when you are determining the popularity of a song. Danceability relates to the overall ability to dance to a song based on key factors such as tempo, rhythm stability, beat strength and overall regularity. This is an important thing to take into consideration when you are attempting to push your song into the market and you are aiming to have a song that can be a hit in a lively environment, such as a club, or other social environment. As far as overall popularity, it was a determining factor in all of the quarters that we broke down and had a positive coefficient throughout the overall dataset as well when determining popularity of songs. Based on the fact that it has a positive coefficient throughout all quarters, it can be concluded that having a higher danceability will result in a song being more popular.

While it's already been noted that the atmosphere of where you're pushing the music can have a varying danceability, geographic factors also come into play. Based on the actual location that your song is coming from as well as being played, the culture behind dancing can determine just how much danceability they would want in their music. Our specific analysis determines what is likely to push the overall popularity though. With tastes constantly changing, it's important to pay attention to the actual interests of those you are pushing the music to.

	Q1	Q2	Q3	Q4	Overall
Std Dev	0.173	0.163	0.176	0.175	0.176
Min	0.059	0.058	0.059	0.055	0.055
Max	0.957	0.974	0.988	0.987	0.988
Average	.5005	0.481	0.543	0.580	0.530

Based on the chart above, we can see the overall average as well as the average of each quarter. Other variables are included such as Std Dev, min, and max but for this variable specifically the average is the main focus of what matters here. This is one of the few variables that maintains a very balanced average throughout the century (see Figure 2). The coefficients give similar information on the results and further prove that maintaining a high danceability is an important factor of a song's popularity. This makes sense when understanding that dance is a physical representation of the art of music.

Apps such as Tik Tok, youtube, and other social media platforms continue to prove to be crucial to the success of a song as well as the shareability of it. With Tik Tok in particular, it’s important to keep in mind that the danceability of a song pushes more shares among the platform as users will start to dance to that song. With trends becoming a big pusher of new artists as well as artist’s music, it’s important to remain up to date with new ways to market songs and how those affect the variables. For example, a higher danceability score will allow for users on other platforms to be able to engage with the music more, resulting in them searching for said music elsewhere. By engaging users through an expression of dance, we can motivate listeners to consistently seek out new music they hear, thereby promoting music and attracting customers to music streaming platforms that highlight these variables in the music they promote.

Speechiness

We selected speechiness as an independent variable because we saw that it had a consistently high negative effect through our data analysis (see Figure 3). Speechiness detects the presence of spoken words in a track. The more speechy a track is the higher the score listed and this is important when we want to recommend how much spoken word an artist will include in their music. An important note about this variable is the limits we set to the actual score itself in our data cleaning process. We further elaborated on this in the data cleaning portion, however, it's important to once again note the fact that our maximum for speechiness is set to a hard limit of .66. This was done on purpose in order to eliminate any sort of audio books or podcasts that would be in the dataset, since spotify does include these things. When determining how we can benefit a song's popularity though we want to eliminate anything that isn't a song. As we continued to go through and familiarize ourselves with the dataset and the variables that we were listing, we went through and confirmed our results as well. The overall max for the dataset was a high .9 and was a german chapter book that was being read with zero music in the background whatsoever. We then listened to something closer to our max, at around a .7 and this was a podcast that had limited sections of music involved in it. As listed in the variable definition, anything around a .3 and above was likely to be rap, hip hop or other fast spoken lyrics.

When creating a song the context of how much speechiness you want to include can vary depending on the genre you are aiming for. It's important to note that our dataset did not provide a listed genre, so when creating music that is intentionally more instrumental the context of the speechiness included can vary based on genre. With a rise in genres such as EDM, Lo-fi, and others with few lyrics, you're going to notice a much lower value of speechiness. As time goes on rap and hip-hop maintain a comfortable position as genre leaders in music and these will have much higher values of speechiness. Another important thing to keep in mind when looking at the speechiness, is what type of activities the user is engaged in at the time. Activities such as studying or reading will want fewer spoken words in their songs. Whereas those on road trips and singing karaoke, will want to listen to music with more words in order to be able to sing along. Another key factor is the listeners mood at the time can affect how much speechiness they want in a song. This is important to keep in mind when referring a playlist to a listener and can allow for slight variations in the allowance of spoken words in your music. By pushing the proper playlists at the right time, based on listener's interests, the streaming platform can assist songs with a variety of speechiness become more popular.

	Q1	Q2	Q3	Q4	Overall
Std Dev	0.088	0.0701	0.064	0.093085	0.337
Min	0.024	0.0232	0.0222	0.223	0.000
Max	0.660	0.6600	0.6590	0.658	0.660
Average	0.085	0.0629	0.0632	0.0891	0.073

Acousticness

Then, acousticness is another independent variable that the group selected to look into more. The reason is because acousticness helps measure the confidence measure between 0 to 1 and whether the track is acoustic. Acoustic is important to look into because it helps the audience hear the same song but in a different format. Many audience members could prefer listening to acoustic music because it gives an authentic voice of the artist. Also, the acoustic form of a song gives the listener a new form of the music from the normal version of a song. Further, it’s important because it would allow the artist to understand if they would like to release a song that includes an acoustic version. While Spotify also could analyze if the customers enjoy acoustic versions too.

	Q1	Q2	Q3	Q4	Overall
Std Dev	0.173	0.7356	0.320	0.309	0.383
Min	0.00005	0.000	0.000	0.000	0.000
Max	0.996	0.996	0.996	0.996	0.996
Average	0.918	0.736	0.338	0.250	0.503

When evaluating the acousticness and how it has changed over the various quarters, it's important to see that the acousticness is often decreasing in each quarter (see Figure 4). It's not to say that acousticness won't be as important anymore in creating a song, but it's something that the artists and Spotify should be mindful of. When an artist is creating a song, they could study these past trends and how acousticness has decreased over time and to take this into consideration. It's to mention that they should be mindful of their approach in writing a song in today's present time. Also, Spotify should be mindful of which songs their customers are listening to when it comes to acousticness. The reason is because Spotify can judge which songs they would potentially like and be able to adjust the recommendations.

Instrumentalness

Instrumentalness was an important independent variable to look into because it allows us to see if a song contains no vocals or not. The vocals in a song are examples consisting of "Ooh", "aah", and more. Additionally, it would be important to look into instrumentalness because there have been studies that people listening to this type of music have experienced lower anxiety, less stress levels, helping with depression, and more. Further, if an artist wants to create a memorable song or if Spotify wants to consider attracting or retaining their audience, then instrumentalness is an important independent variable to consider.

	Q1	Q2	Q3	Q4	Overall
Std Dev	0.394	0.346	0.267	0.335	0.338
Min	0.000	0.000	0.000	0.000	0.000

Max	0.994	0.994	0.999	1.000	1.000
Average	0.394	0.217	0.124	0.191	0.204

For instrumentalness, there are various changes throughout the quarters. In quarter one, it shows that instrumentalness is a lot higher in comparison to the other quarters (see Figure 5). However, moving forward to the middle two quarters, such as quarter two, it drastically drops compared to the others. Lastly, in the current time period, the instrumentalness then drastically increases compared to the other variables. Instrumentalness can change quickly through many decades and an artist should be mindful of this in what decade they are releasing their songs. Or, Spotify can evaluate their customers and see how they feel about instrumentalness in their playlist and evaluate that as well. All in all, instrumentalness fluctuates a lot throughout the decades and isn’t a steady pattern.

Results

Table of Coefficients					
Attribute	**Q1**	**Q2**	**Q3**	**Q4**	**OVR**
Intercept	**2.4356**	**30.5857**	**42.3996**	**74.0602**	**53.4821**
Acousticness	**-1.0692**	**-20.5412**	**-0.9403**	**4.9915**	**11.9727**
Danceability	**3.884**	**9.7553**	**6.4551**	**-3.2289**	**1.7978**
Instrumentalness	**-1.5132**	**-5.9923**	**-2.2311**	**-24.4012**	**-27.3473**
Speechiness	**-5.2365**	**-31.7289**	**-7.5326**	**-11.3008**	**-15.0596**

The table of coefficients show the song attribute interests for songs released in the past 100 years. While this table only shows the four coefficients across each quarter. Each quarter was run against all the other significant variables in that quarter. Since there were more significant variables in each, it was helpful to the main four's effect on popularity when compared to the other significant variables as even the most influential variables explained a small portion of the model (looking at the max R^2 value of 25.6 percent). When looking at the first quarter, we can see that the coefficients tend to hover around 0 - or at least much more so than the other quarters (see Figure 6). From this, we can infer that there isn't a strong overall preference for a specific type of song in this quarter. A strong preference for a certain attribute would look like the coefficient being noticeably high or low (for example, speechiness in quarter two being -31.7289). In quarter one, popularity will generally increase by 3.884 for an increase in danceability by one, but there is still a preference or strong tolerance for the other variables. Listeners tend to enjoy speechiness less in songs of this era (popularity decreases by -5.2365 for every increase by one), but overall, its impact is not significantly negative.

Moving to quarter two, we see a massive change in listener preferences (see Figure 7). People have a much stronger preference for what attributes make up a song - the lower the value in acousticness (-20.5412) and speechiness (-31.7289), the better. There's a significantly lower tolerance for each attribute in this era. In quarter one, the popularity of a song didn't overly depend on a specific attribute, but quarter two, there is a definite template.

Quarter three is similar to quarter one in that there is much more of a tolerance of the different variables as they hover closer to zero (see Figure 8). If we look at the type of music during this era, the 70s, 80s, and early 90s saw a great influx of new music from rock, metal, disco, pop, and electronic music. Each genre had their own songs that were quite popular and naturally each genre has their own unique mix of attribute levels. For example, both rock and disco songs could be equally popular despite the fact that disco songs might have higher levels of danceability and instrumentalness. Rock could make up the difference by having low levels of acousticness and speechiness. Listeners slightly prefer more danceability (6.4551) in a song which aligns with the genres of the time.

Lastly, it's important to notice how the most recent quarter encourages songs to be a specific type of song across almost all categories due to their negative coefficients (see Figure 9). A strong decrease in instrumentalness (-24.4012) could be explained by the rise in popular rap artists over the last two and half decades as rap songs are mostly lyrically based. The overall trend looks similar to quarter four as 29 out of the top 30 songs are from the most recent quarter (see Figure 10). The overall trends should be similar to that of quarter four and since on average, instrumentalness (-27.3473) and speechiness (-150596) have a large negative effect across the quarters, those effects would be quite large on the overall popularity.

With the overall look at trends over time, we decided that quarter four is the most important time period to focus on in predicting song popularity. As stated earlier, 29 out of the top 30 songs are from the last 25 years and since quarter four's personal preference in song type is very different than the other quarters (low instrumentalness at -24.4012 and speechiness at-11.3008), this would be the best factor in anticipating what would make a new release popular among current listeners. The low instrumentalness and speechiness of this quarter can likely be attributed to the rise of popular rap artists such as Drake, Travis Scott, and Eminem as rap is very much lyrically based compared to other genres. There were other more influential coefficients such as explicitness however, we anticipate the reason why is not because the more explicit songs are, the more popular they become, but rather explicit songs get much more exposure in today's society. Back in the day, one of the key determinants of song popularity was whether it received significant radio play time. To be played on the radio, a song couldn't be explicit or else it wouldn't get that play time. Nowadays, with streaming services such as Spotify and YouTube, those restrictions have been lowered so explicit songs can suddenly be found by millions of people online. As this is the case, we anticipate that this variable is a case of correlation instead of causation.

If we then run a regression with just those four variables instead of comparing them to the other variables in the model (all that weren't significant across all quarters), then we can evaluate the coefficients as an equation to predict popularity. The equation would be as follows: $\square = 33.821 + 7.237\square_1 + 10.293\square_2 - 33.375\square_3 + 4.256\square_4$.

In this case, $\square_1$ = acousticness, $\square_2$= danceability, $\square_3$= instrumentalness, $\square_4$= speechiness. We can see that for an increase in acousticness to .996 (the highest it can go), popularity would increase by 7.21. Additionally, if we have a song with a danceability score of .988, popularity would increase by 10.169. This logic can be applied to the other variables in the equation. In theory we could put the attributes of a song and try to predict the exact popularity of a song but as we stated earlier, with the R^2 being so low, it wouldn't be an incredibly accurate prediction. This equation just helps visualize how the coefficients affect popularity.

In an attempt to prove the results of our analysis, we chose to select the song with the highest popularity score in the dataset and analyze the values for each of our selected variables. This song was 'Driver's License' by Olivia Rodrigo. As seen in Figure 11 in the appendix, the outcome of our analysis yielded the following values; Danceablility: 0.585, Speechiness: 0.0601, Acousticness: 0.721, Instrumentalness: 0.0000131. In comparing this to the results of the overall values from the fourth quarter; Danceablility: 0.58, Speechiness: 0.0891, Acousticness: 0.25, Instrumentalness: 0.191. We can see that 'drivers license' maximized the variables with positive effect on popularity such as, around triple the acousticness of the average, and minimized variables with negative effects such as, almost zero instrumentalnss and a 43% decrease in speechiness compared to the average. We believe that this comparison reinforces our findings as the most popular song in the dataset has maximized values that we have found to have a positive effect on popularity and minimized those that have a negative effect.

Business Applicability

We approached this problem with a multifaceted approach, we aimed to cater to both artists and listeners, recognizing the mutual benefit in doing so. By identifying key variables impacting a song's popularity, we sought to inform artists about what resonates with listeners. Encouraging artists to create music with attributes found in popular songs from our analysis as including these attributes can enhance their chances of current success. On the listener's side, leveraging listener information, we assign scores to different attributes for each listener, aligning their preferences with songs with those attributes to facilitate better matches between songs and listeners. We plan to implement this via an option playlist listed on the user's home page. Similar to the "Because you watched this," feature in Netflix. Offering users the opportunity to explore a tailored playlist of new music, rather than pushing it onto the user. Through this innovative feature, we enhance the music discovery experience for both new and existing users, thereby attracting more customers to our platform. Simultaneously, by providing exposure to new audiences, we support artists in gaining traction and recognition for their work. This dual approach, focusing on providing value to both artists and listeners, ultimately serves to strengthen the company as both customer bases derive greater value from Spotify's service, fostering continued growth and success.

Appendix

Figure 1:

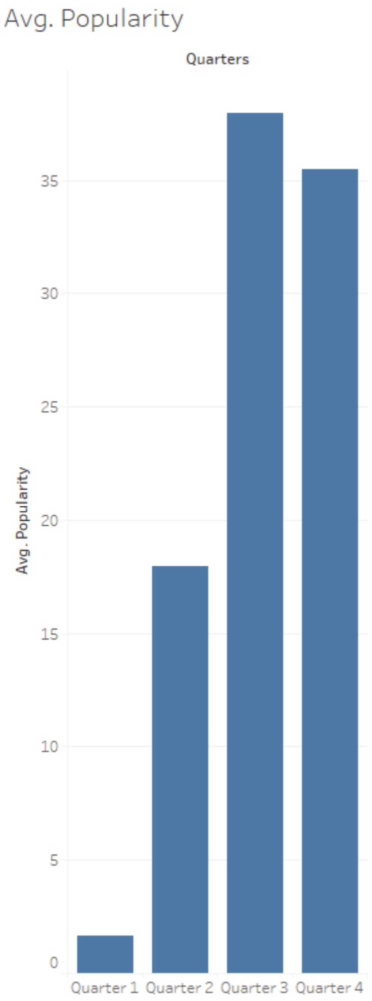


Figure 2:

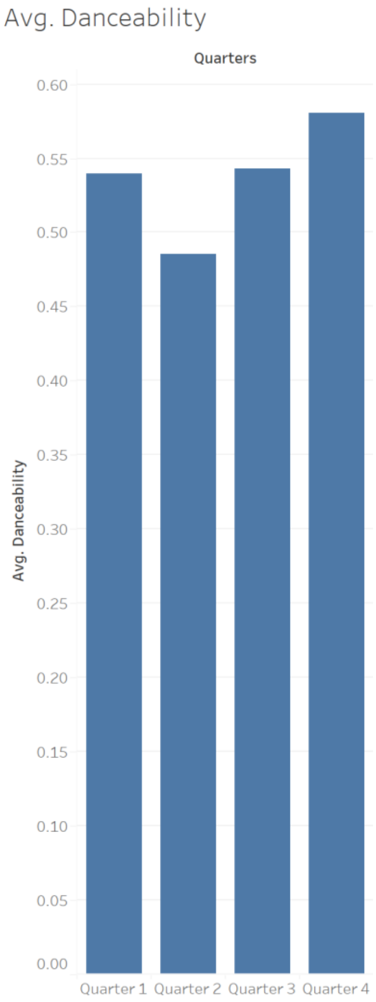


Figure 3:

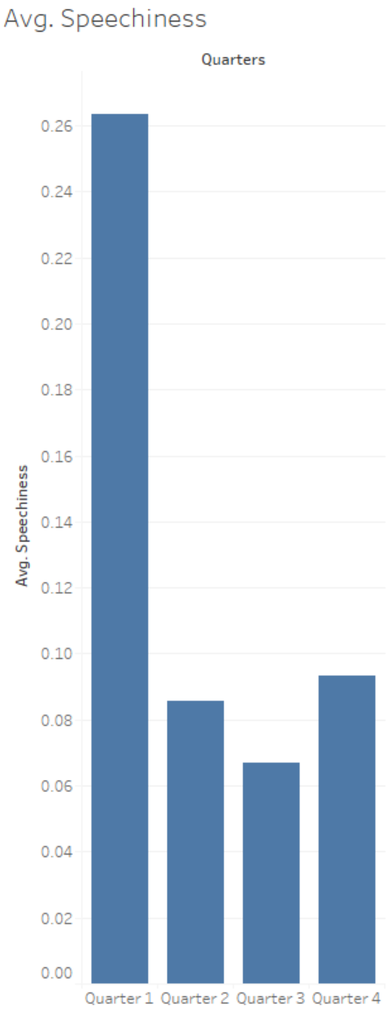


Figure 4:

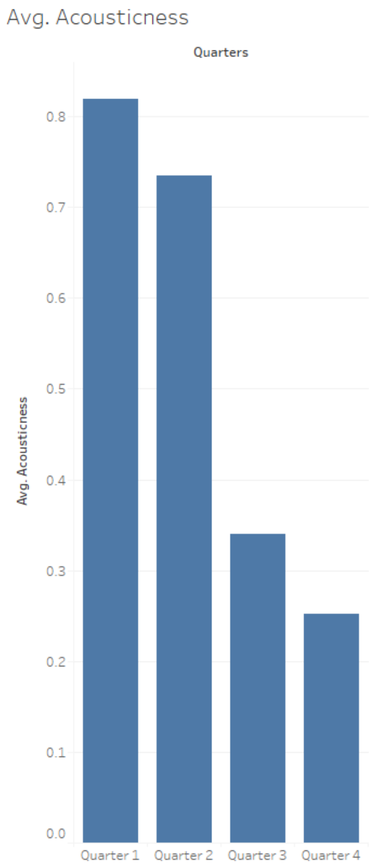


Figure 5:

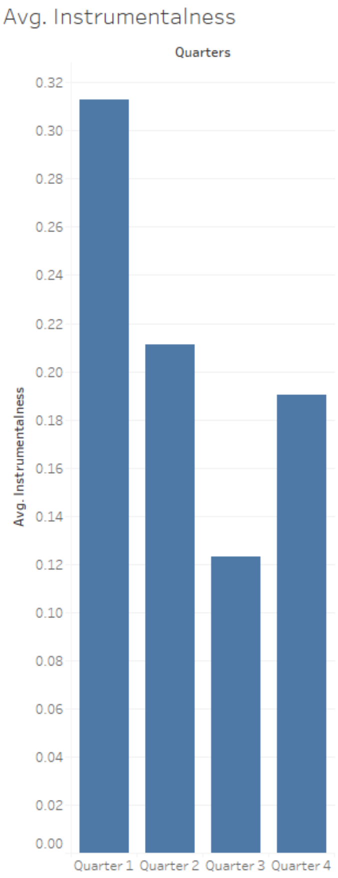


Figure 6:

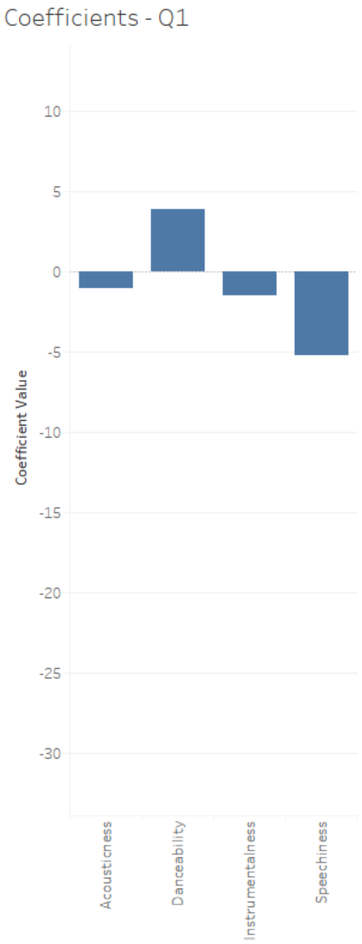


Figure 7:

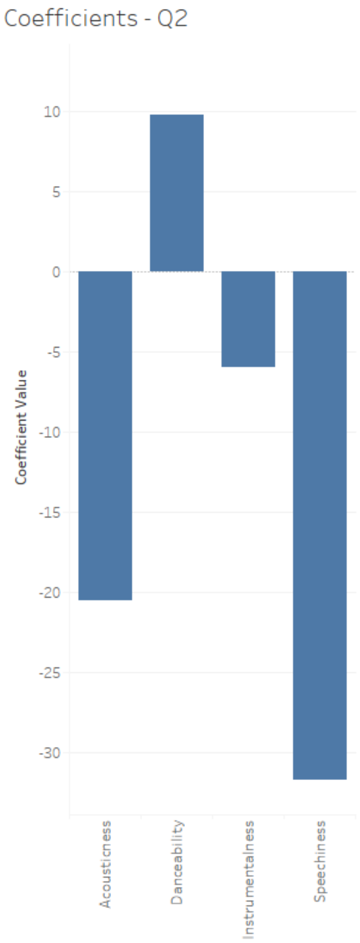


Figure 8:

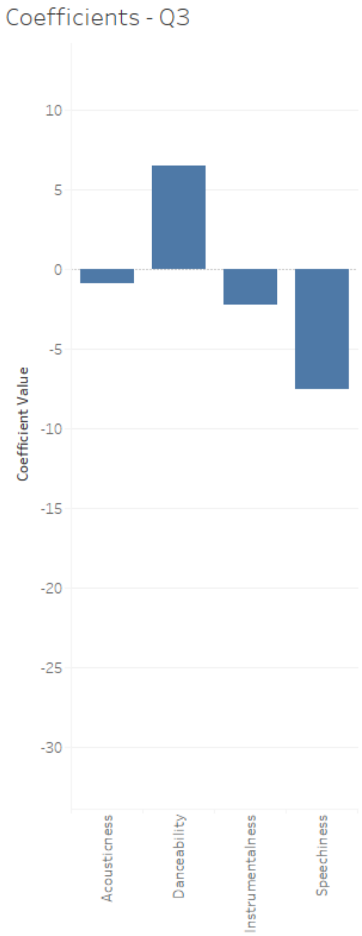


Figure 9:

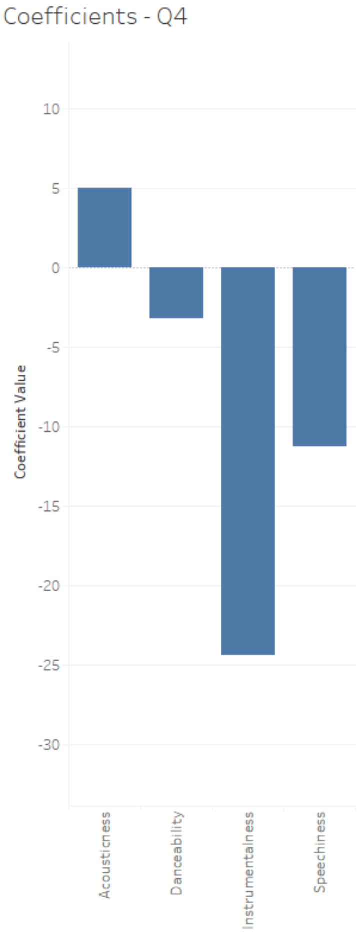


Figure 10:

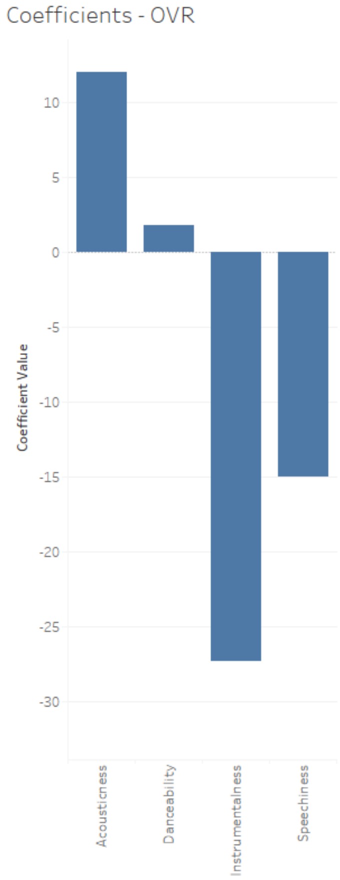
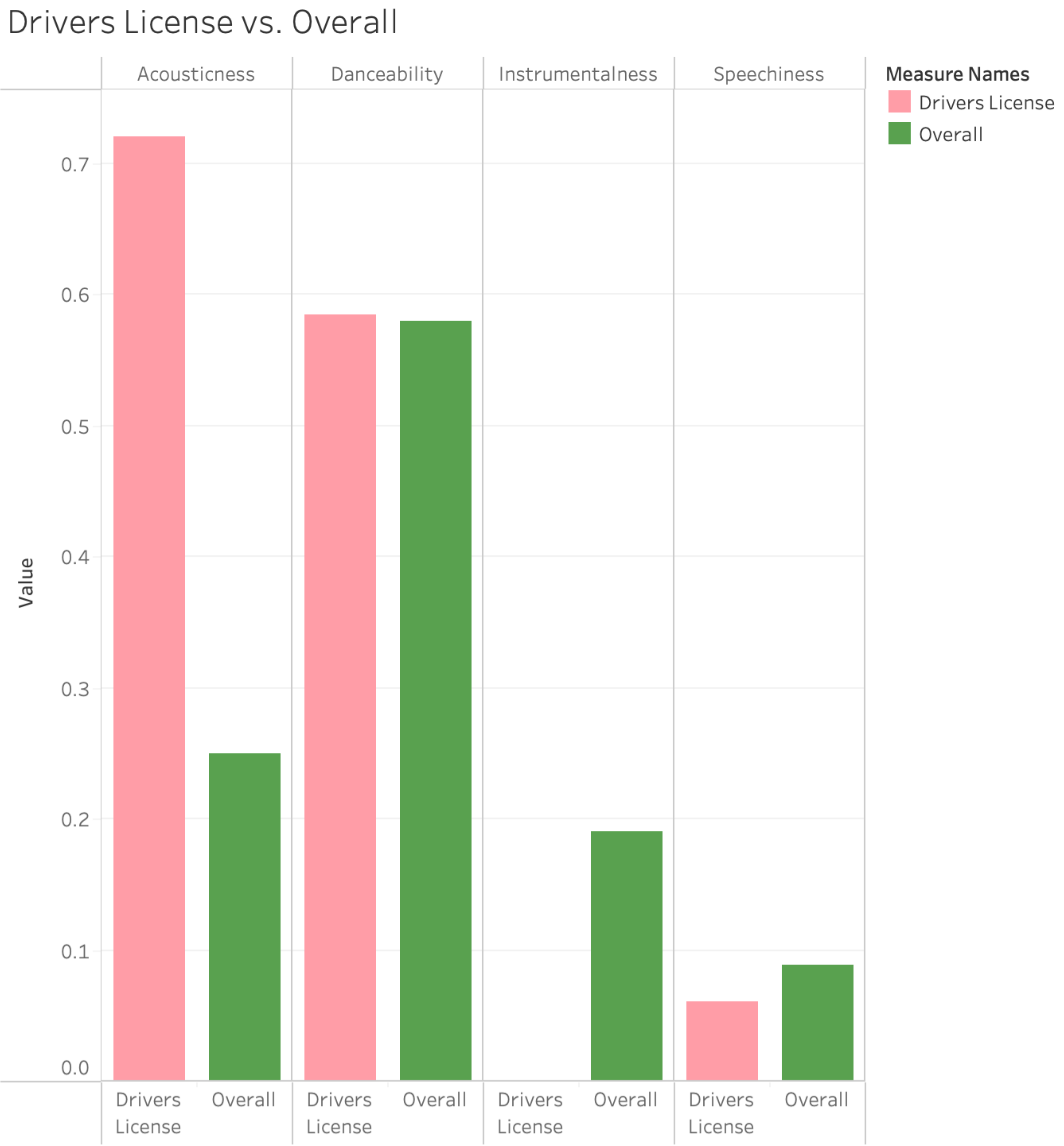

Coefficients - OVR
Coefficient Value
10
5
0
-5
-10
-15
-20
-25
-30
Acousticness
Danceability
Instrumentalness
Speechiness

Figure 11:


Drivers License vs. Overall
Acousticness
Danceability
Instrumentalness
Speechiness
Measure Names
🟥 Drivers License
🟩 Overall
Value
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Drivers License
Overall
Drivers License
Overall
Drivers License
Overall
Drivers License
Overall