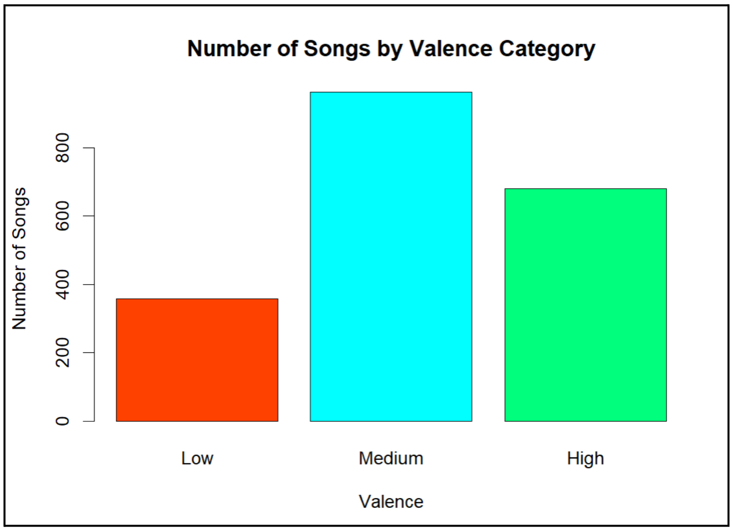
This is how you make the Most Popular Song...

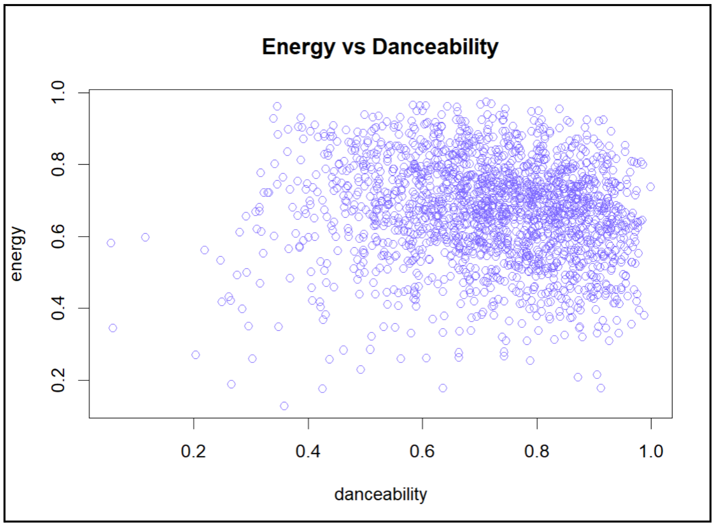
Annika Bhatia **-** Data 101 - Dr. Imelinksi - 9/24/24

Music is one of the things that constantly changes with time. As time goes on, so do music trends. Listeners can hear the differences in the most popular artists and genres. So, the question remains, what makes the most popular song?

I conducted an analysis on a Spotify Dataset that displayed the Top Hits from 2000-2019. In this dataset, there were many columns, such as danceability, acousticness, duration in milliseconds, genre, valence, and much more. When looking at this data, I started to notice that a majority of the categories given had an interesting association with the popularity of the song. So, I decided to analyze what exact factors helped make a song popular over the past two decades.



I first analyzed the valence of all the songs in the dataset. The valence of a song is the measure of musical positiveness conveyed by a track. I first categorized the valence into Low, Medium, and High. The low values were from 0-0.33, medium values were from 0.33-0.66, and the high values were from 0.66-1. As you can see, the songs with lower emotional tone had the lowest number of songs, then the higher emotional tone ones, and then the medium ones had the highest. This was interesting to me since I expected the high category to have the largest number of songs. This graph helps show that valence is an important aspect to consider when trying to make the most popular song.



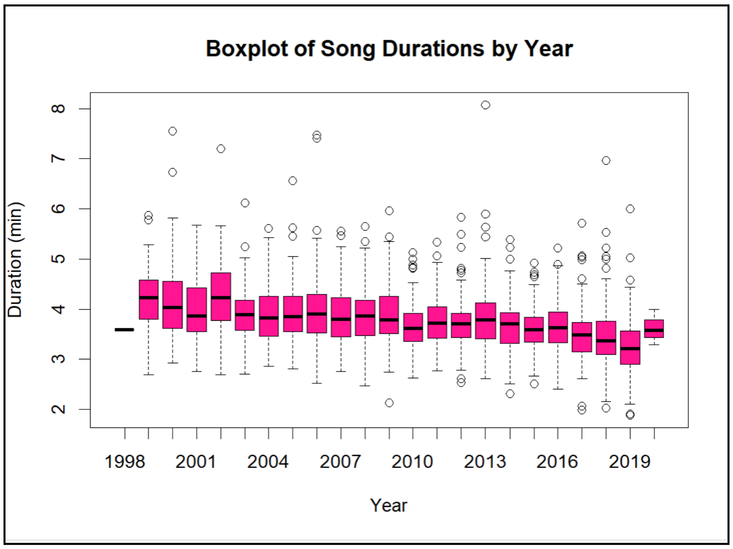
Then I wanted to analyze the energy and danceability of the popular songs. At first, I wanted to see the relationship between energy and danceability. I predicted that there would be a direct relationship between both variables: and I was proven correct. In the scatter plot to the left, the majority of the popular songs had a higher energy level and higher danceability. It's clear that people were looking for songs that were fun and upbeat.

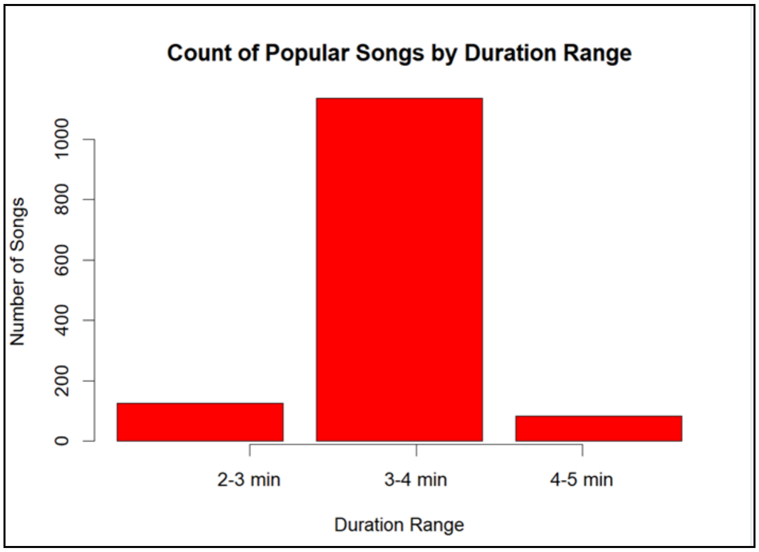
The average duration for a popular song is 3.8 minutes...

***Annika Bhatia* - Data 101 - Dr. Imelinksi - 9/24/24**

One of the prime aspects of a song is the duration of it. Some people prefer shorter songs because they like to constantly switch the songs they are listening to, while others prefer listening to longer songs and really appreciating the lyrics, melody, and tune.

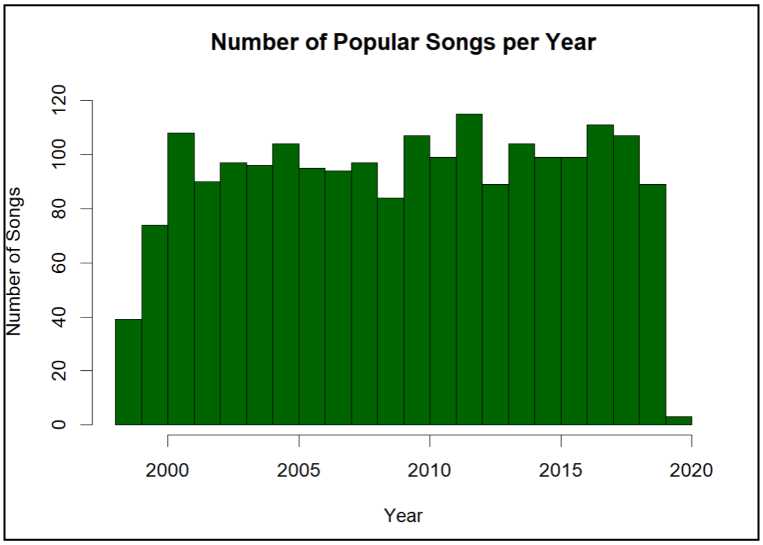
According to my calculations from the dataset, the median of all the popular song durations seemed to relatively stay the same (around three and a half minutes). The mean of the average song distribution for songs with a popularity level of above 60 was 3.802408. The original dataset had the duration in milliseconds, however I created a new column and converted it into minutes so it would be easier to understand. In the most recent years though, the range of the duration of songs started to shrink. This can be correlated with the fact that the human population's attention spans have been decreasing, which could mean that they do not thoroughly enjoy listening to up to four minutes of music. They would most likely rather prefer a song that is on the shorter side (around two and a half minutes).





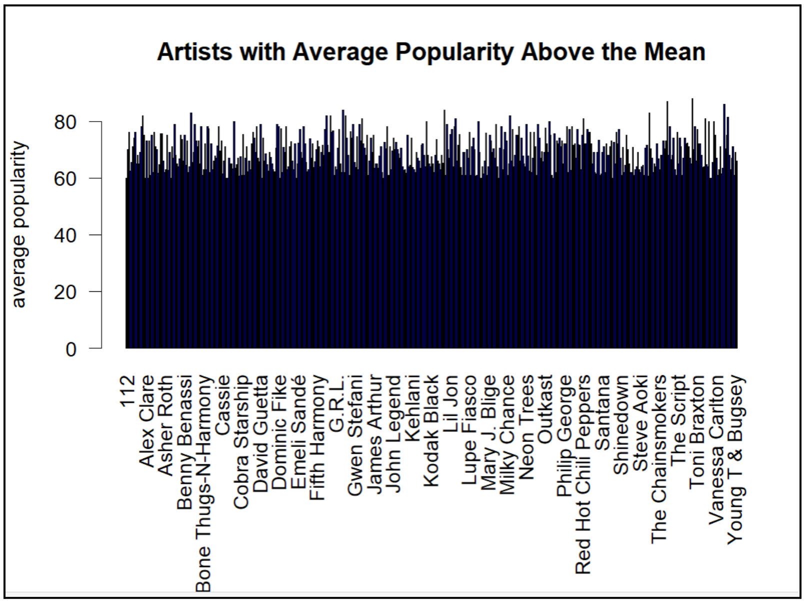
So far from all the data I have gathered, these are the following characteristics that make a popular song: a high valence, high energy, high danceability, around a 3-4 minute duration.

This is the most popular artist...



Another aspect of songs is whether they are explicit or not explicit. On average, the popularity level of songs that are not explicit is 59.3, while the popularity level of songs that are explicit are 61.5. This indicates that songs with more mature language or themes tended to be more popular in the 2000s and up till 2019.

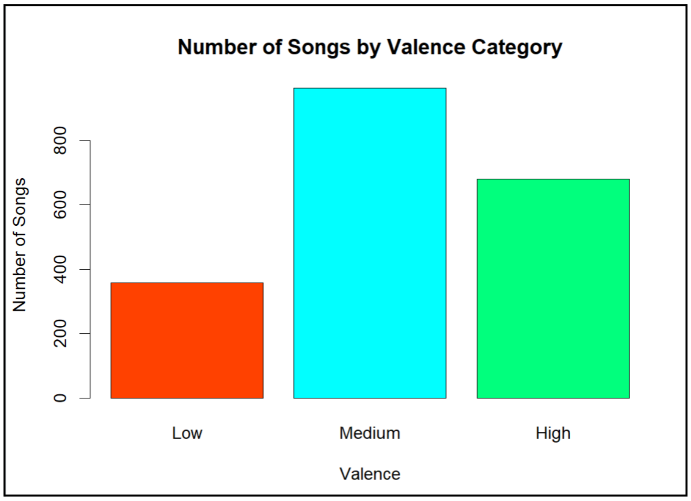
I also wanted to analyze the overall number of popular songs per year. What was interesting to me was that one of the years that had the lowest number of popular songs was 2009, when I believed that to be one of the prime years of music. Artists such as Black Eyed Peas, Florida, and Mariah Carey were very popular at the time, so I expected the value for the 2009 bar to be higher.



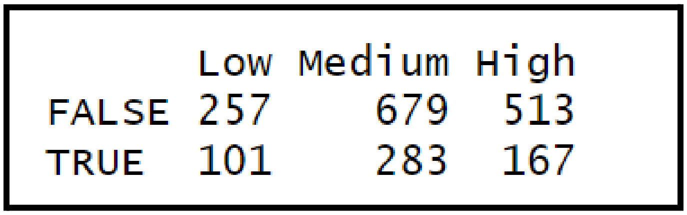
This interesting pattern pique my curiosity on who the most popular artists were, so I decided to create a bar chart of this query. Since there were many, many artists in the dataset, I had to filter the artists column based on the mean popularity. Artists who had a value that exceeded the mean popularity were included in the chart, and those below were excluded. Based on the chart, we can see that the most popular artists (both have very similar average popularity values) were Toni Braxton and Vanessa Carlton. This was very interesting to me because I expected artists like Britney Spears, Pharrel Williams, and artists from more recent to be on the list. It was interesting to see the most popular artist be someone who I am not familiar with.

The odds that a song is explicit given that it is the “Medium” valence category is 0.41…

Annika Bhatia **-** Data 101 - Dr. Imelinksi - 11/12/24



After looking through all of the different graphs constructed for this dataset, I decided to use valence and explicitness as my two categorical variables of the Bayesian analysis. My question I analyzed was: What are the odds that a song is explicit given its in the medium valence category? The observation was that there are songs that fall within the medium valence category. The belief was that the songs in this valence category are explicit.



```
      Low Medium High
FALSE 257    679  513
TRUE  101    283  167
```

The prior odds did not take explicitness into account, and was a value of 0.38. The true positive was that a song that is explicit in the "medium" valence category, and ending up being a value of 0.29. The false positive was that the song that is not explicit but is in the "medium" valence category, and ended up being a value of 0.71. Based on this data, the initial likelihood ratio was 0.41. The posterior odds gave the new likelihood after taking into account new evidence, and was a value of 0.1558. Therefore, to find k, we divided the posterior by the prior to get a value of 0.41 as our *k*. This means that after considering the dataset that led to the posterior odds, the likelihood of the hypothesis being true increased by a factor of 0.41 compared to the prior odds. I also created a contingency table, as shown to the right, with the rows as the belief and the columns as the observations. I got the counts for T[Explicit, Medium] and T[Not Explicit, Low] as the T[i, j] and T[k, l], and got a likelihood ratio of 2.89.