CS5830 Project 3 Report

Gavin Murdock and Nithya Alavala

Introduction

The intent of this report is to give record labels, bands, and individual artists a fresh perspective on popular music and to provide insights on how the music industry might be changing. Our analysis is based on real-world data obtained from one of the top music streaming services in the world, Spotify. Throughout our analysis, we used a variety of statistics and charts to gather and visualize information we think is valuable. This report will discuss a number of these findings along with how they might be valuable to artists and their record labels. The links to our Github project folder and presentation slides can be found below.

Github project, presentation slides.

Dataset

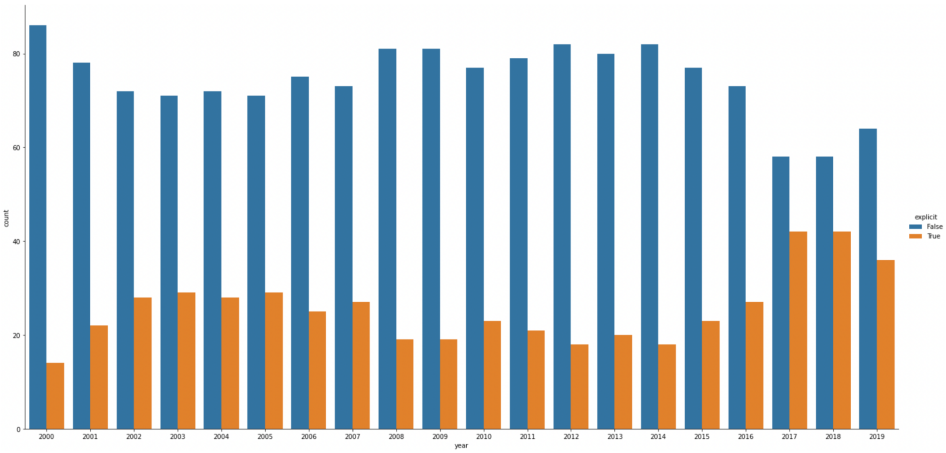
Before starting our analysis, we created our own dataset by gathering song data through the Spotify API. Spotify has 19 official playlists titled “Top Hits of” followed by the years 2000 to 2019, and each of these playlists contains the top 100 songs of that year. The Spotify API has a get playlist tracks endpoint that takes a playlist ID and returns various data about each song in the playlist. We collected the playlist IDs for all of the top hits playlists and used this endpoint to retrieve data we then stored in a csv file to be used as our dataset. We were able to gather a variety of data on each song such as the name of the song, what year the song was a top 100 hit, the artist who made the song, whether the song was a single or on an album, and more. Creating this dataset enabled us to study possible trends and correlations in the music industry over the course of 20 years.

Analysis Technique

We have used a Kernel Density Estimation(KDE) plot for plotting the sheer share of the album type with the popularity each track has, because the KDE is a good way to show the density, or percentages when they add up to 1. We have used a catplot for plotting the number of songs that become explicit over the years and also we have searched in some articles for reasons why the songs are explicit, attribute explicit is boolean here. For seeing a correlation between the duration and popularity, we have used a scatterplot with a regression line,to easily see a correlation between the values, if any. For plotting the duration and number of tracks, we have used a regplot for the same reason. We have used scatterplots to understand the popularity of songs with respect to duration and also songs that are listed in top 100 rank with respect to duration. For our final analysis, we have used a barchart for plotting the percentage of top 100 songs that are singles each year.

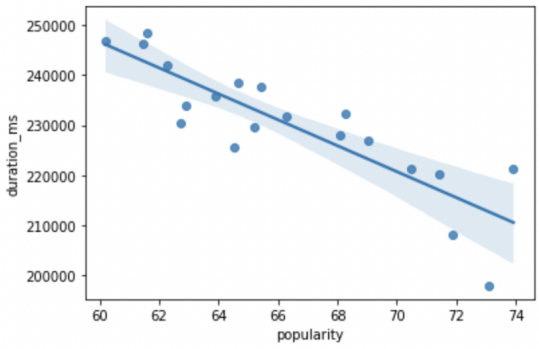
Results

Explicit songs over the years



The catplot depicts how songs become explicit over the years from 2000 to 2019 based on spotify API. According to spotify, explicit songs are offered how the artist intends it to be heard. Explicit label applies when the lyrics or content of a song contain strong language, offensive words or culturally/socially abusive content. An assumption can be made from the chart that in the early 2000's music started to have more explicit content which might have led to this fact from a CNN.com article, citing the meteoric decline of sales during a decade span. Between 1999 and 2009, music sales went from $14.6 billion, to $6.3 billion; a decline of over half. Another assumption can be made that in the year 2017, more explicit songs as shown in the chart were made such as Despacito, albums from Calvin Harris,Drake,Jay-Z etc. It was also mentioned in an article that 75% of songs in the top 100 were explicit in 2017 and it's growing since then. In the year 2018, the count was the same as 2017 since songs such as IDGAF from Dua Lipa, FRIENDS from Marshmello etc had explicit lyrics based on spotify.

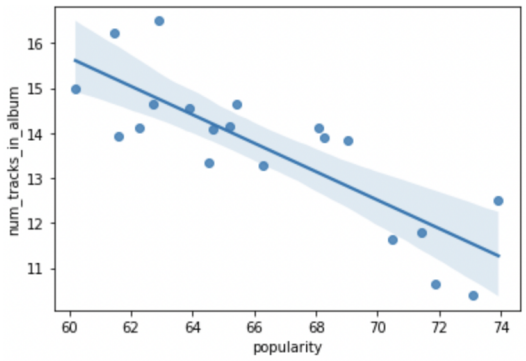
Duration vs Popularity Correlation



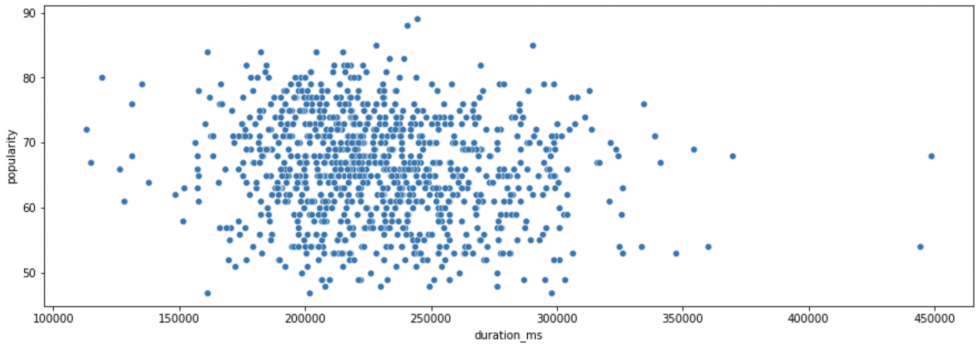
To understand the correlation between popularity and duration of a track, we applied the Pearson correlation coefficient to measure the correlation, and we got r and p around -0.85 and 1.22e-6 respectively which means that they are strongly negatively correlated. Changes in one variable are correlated with changes in other variables. From the regression plot, we can say that if the duration of a track decreases, the popularity increases.

Duration vs Number of tracks Correlation

From the regression plot, we can say that there appears to be a strong negative correlation between popularity and number of tracks in an album. We calculated the Pearson correlation coefficient and we got r and p values as -0.82 and 9.12e-6 respectively. We can confidently say that there is a strong negative correlation. If the number of tracks in an album increases,popularity is more likely to decrease since there's an assumption that people might not have time to listen to all the songs in an album unless they are popular and that all the songs might not grab people's attention.

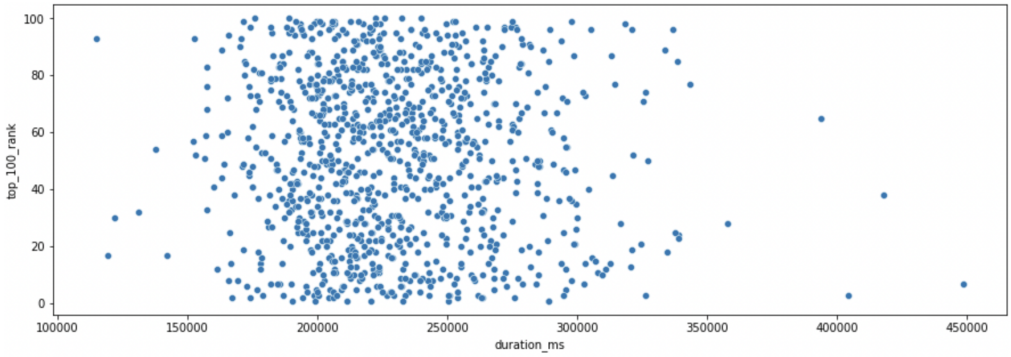


Duration vs Popularity



This scatterplot shows that when the duration is less than 2.5 minutes, there are only a few songs that are popular and the number of songs with duration between 3 minutes i.e, 180000ms and 4.5 minutes i.e, 270000 ms are more in count compared to the ones having more duration. We can assume that people might not listen to songs that are longer than 5 minutes. We can say that tracks with duration less than 3 or more than 5 minutes are very few that became popular when compared to the ones that lie in between, there are many songs that became popular according to the plot.

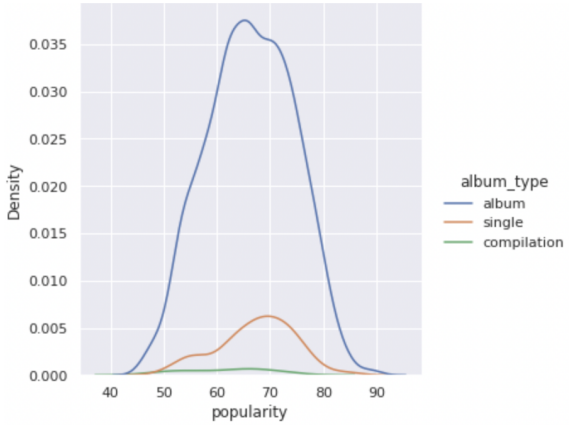
Duration vs Rank



Similar to the previous one, this scatterplot shows that when the duration of a song is less than 2.5 minutes, only a few songs entered the top 100 list while there are many songs that got listed

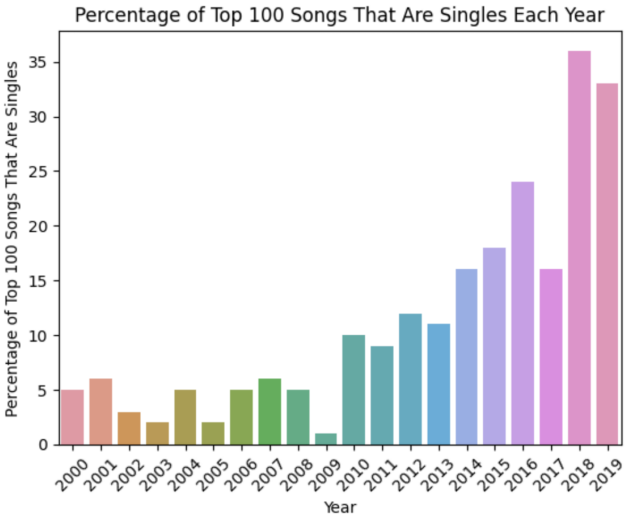
have the duration more than 3 minutes. And the ones with more than 5 minutes are only a few that are in the top 100 according to spotify.

Popularity vs Album type



We ran a hypothesis T-test considering the popularity for types of albums as parameters. From the data we grouped the types of albums to get the number of songs of each type and check which album type became popular. We then noticed that most of the songs are of album type, 'Album'. We performed the T-test on album popularity and compilation popularity and got t-statistic and p-value around 3.189 and 0.0014 respectively which says that they are not similar. Another test on album and single popularity with -1.859 t-statistic and p value of 0.0631 represents that they are similar. For single and compilation popularity, t-statistic was around 4.134 and p value 4.843e-5 which says that they are different. Our analysis will be helpful for bands, artists to know how the music industry is evolving.

The Rise in Singles



As this bar chart illustrates, there has been a significant increase in the percentage of top 100 songs each year that are singles. In fact, a Pearson correlation test between year and the percentage of the top 100 songs each year that are singles produced an r value of 0.8 and a p value of .000004, showing there is a strong and significant correlation between the two. With this rise in top 100 singles, record labels and musicians should be pushing to release more pre-album singles in hopes of creating a hit.

Technical

For data collection, we used Spotify's API. We subscribed to the free version of the API through RapidAPI, where we retrieved an API key. We then used Python's requests library and for loops to send GET requests to the API's playlist tracks endpoint, parse the returned JSON, and

convert the data to a csv file in order to make future access of the data easier and more accessible.

For analyzing how songs become explicit over the years, we used a catplot from seaborn to plot the number of songs which are explicit.

For duration and popularity analysis, we grouped our data by year and then aggregated other attributes such as popularity, duration and number of tracks in the album. We then applied the Pearson correlation coefficient to see the relationship between popularity and duration. We can say that if the duration of a track decreases, the popularity increases.

For duration and number of tracks in the album, we performed the same kind of analysis as we did for the previous analysis, we looked for the correlation between these variables. We can say that if the number of tracks in an album increases,popularity decreases.

To understand how the duration of a track affected popularity, we plotted a scatter plot to visualize the data and see how many songs became popular when length of a song is considered. We did a similar analysis to understand how the duration of a track affected rank.

For popularity and album type analysis, we used a Kernel Density Estimation(KDE) to plot different types of albums with popularity. And then we performed a hypothesis t-test on these types of albums.

For the rise in singles analysis, we grouped our data by year and then aggregated attributes to get singles count. We have used a barplot to see how the percentage of top 100 songs increased each year. We then applied the Pearson correlation coefficient to see the relationship between year and the percentage of top 100 songs that are singles and we observed a strong and significant correlation.