Investigating Popularity of Songs on Spotify

Dancing Queens: Karianna Klassen, Madeline Waterfield, Mark Yukelis, Shrikrishna Sriram

Introduction and Data

Background and Significance

Spotify, a Swedish company founded in 2006 by Daniel Ek and Martin Lorentzon, is self-described as "the world's most popular audio streaming subscription service." Across 183 markets, Spotify has 406 million users, 180 million subscribers, 82 million tracks, and 3.6 million podcasts. Unfortunately for artists, the large number of users does not change the fact that only a small percentage of the 82 million tracks reach over 1 million streams. According to Spotify's Loud&Clear website intended to increase transparency for artists, earning 1 million streams would put a song in the top 719,000 tracks, which means that only 0.87% of songs recieve over a million streams. Only 240 songs have reached Spotify's "Billions Club" by earning over one billion streams. For small artists, "going viral" or earning a significant number of streams is difficult and unlikely. In an effort to help artists understand how to create a song that will be succesful on Spotify, we will investigate attributes of popular songs and endeavour to reach a consensus on what makes a song more successful than others.

Data

In the following project, we will investigate a dataset of 30,385 Spotify songs in order to determine whether songs with certain characteristics are more likely to be popular than others. The dataset spotify_songs.csv is from a TidyTuesday launch on January 21, 2020. The data was gathered by spotifyr, an R wrapper for pulling track audio features and other information from Spotify's Web API in bulk. The original dataset has 32,833 observations, or songs, and 23 variables, which means that for each song there are 23 variables to identify it, like song name, artist, tempo, playlist name, and release date. However, because some songs are put on multiple playlists, those 32,833 observations include some duplicate songs. After cleaning the data by removing some playlist identifiers and keeping only distinct observations, our dataset has 30,385 observations, and each observation is a song. In cleaning the data we had to remove three variables associated with playlist information (playlist name, playlist subgenre, and playlist ID), so 20 variables remain. Thus, the dataset we will be working with has 30,385 observations and 20 variables. We will be focusing on five variables: popularity, genre, danceability, speechines, mode, and duration. Song popularity refers to the popularity of a song (or number of streams) relative to other songs in the dataset, ranked from 0-100 (where higher is better). Genre refers to the genre of the playlist the song is located on, of which there are six: Pop, Rap, R&B, Latin, EDM, and Rock. According to TidyTuesday, "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." Mode describes whether the song is in major or minor key (major is represented by 1 and minor is represented by 0). TidyTuesday describes speechiness as "the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks." Song duration refers to the length of the genre in minutes. Popularity, genre, danceability, mode, speechiness, and duration were originally determined by Spotify's Web API.

Research Question

Our research question is: **What characteristics make a song more likely to be popular than other songs?** We are endeavouring to find out what qualities make a song popular in order to be able to tell artists what kind of song has the most likelihood of becoming popular. We can break up our research question into five smaller research questions:

Research Question 1: Do some genres have more popular songs than others? Before analysis, our hypothesis is that pop music and rap music will have the greatest proportion of popular songs.

Research Question 2: Do more popular songs have a significantly different danceability than less popular songs? Before analysis, our hypothesis is that more popular songs have a higher danceability than less popular songs.

Research Question 3: Are more popular songs signficantly more likely to be in the major key than less popular songs? Before analysis, our hypothesis is that more popular songs are more likely to be in major key than less popular songs.

Research Question 4: Do more popular songs have a significantly different speechiness than less popular songs? Before analysis, our hypothesis is that more popular songs and less popular songs have the same average speechiness.

Research Question 5: Do more popular songs have a significantly different duration than less popular songs? Before analysis, our hypothesis is that more popular songs have a lower duration than less popular songs.

Methodology

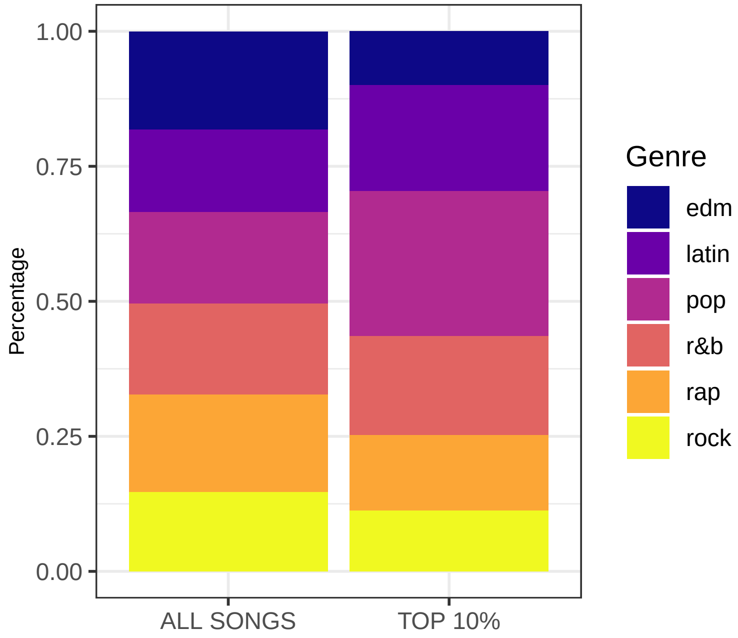
When conducting our analysis, our team sought to understand what qualities make a song more likely to be popular. To do this, we decided to compare the mean attributes of all of the songs on the list with the mean attributes of the top 10% of songs in terms of popularity, hoping that differences (or continuities) in summary statistics would give us insight into our research questions.

Exploratory Data Analysis

Genre

Figure 1: Proportions of Genre

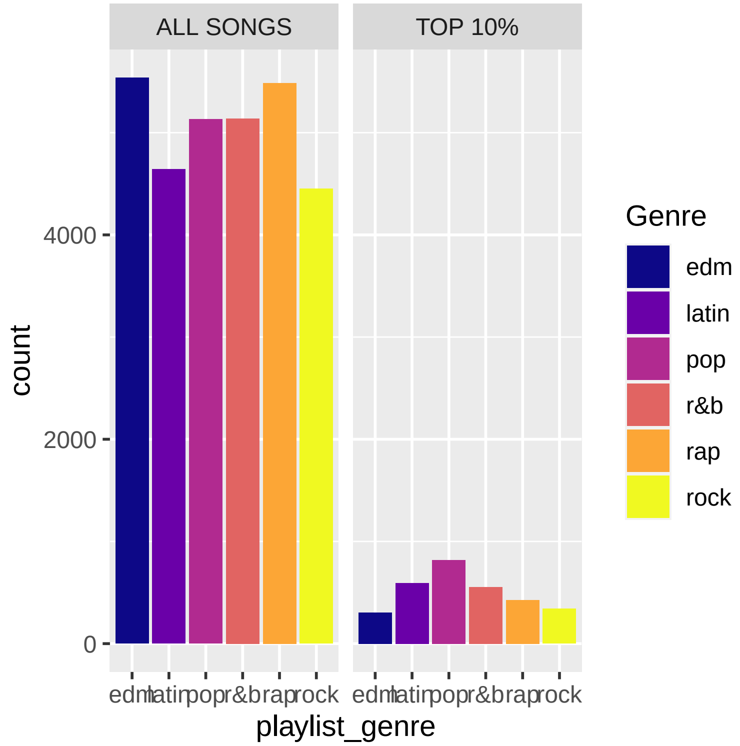
For All Songs vs. the Top 10%



As seen in Figure 1A, there are about the same amount of songs from each of the six genres: 5512 EDM songs, 4641 Latin songs, 5175 Pop songs, 5486 Rap songs, 5120 R&B songs, and 4451 Rock songs. However, as seen in Figure 1B, the proportions change for the to 10% of songs. In the top 10% of songs there 304 EDM songs, 594 Latin songs, 818 Pop songs, 426 Rap songs, 555 R&B songs, and 341 Rock songs. This means the proportions change from 18.14% EDM, 15.27% Latin, 17.03% Pop, 18.05% Rap, 16.85% R&B, and 14.65% Rock in the whole dataset to 10.00% EDM, 19.55% Latin, 26.93% Pop, 14.02% Rap, 18.27% R&B, and 11.22% Rock in the top 10% of songs in terms of popularity. The largest increase in proportion occurs for Pop songs: the proportion of Pop songs is 10.08% higher for the top 10% of songs when compared to the overall proportions.

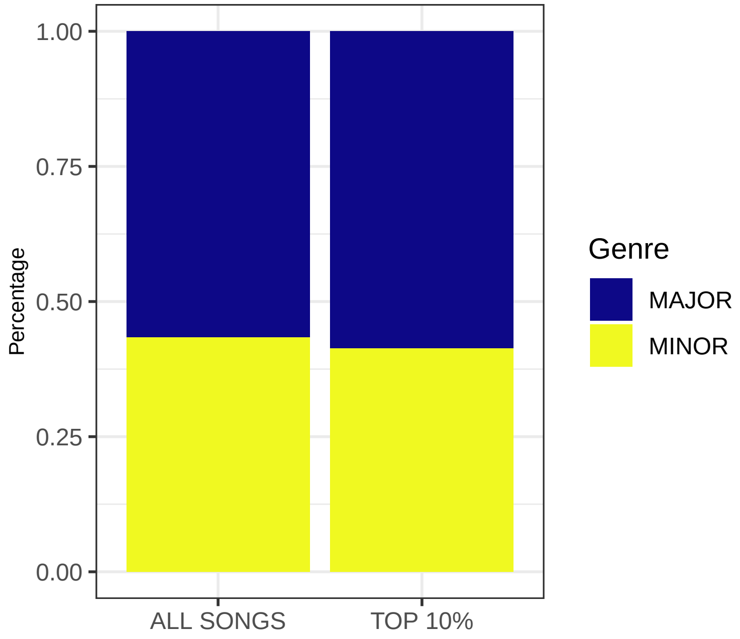
```
## # A tibble: 1 x 1
##   `mean(track_popularity)`
##                      <dbl>
## 1                     45.9
```

```
## # A tibble: 1 x 1
##   `mean(track_popularity)`
##                      <dbl>
## 1                     41.0
```



Mode

Figure 2: Proportions of Songs in Major Keys
For All Songs vs. the Top 10%


Percentage
ALL SONGS
TOP 10%
Genre
MAJOR
MINOR