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 receive 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.

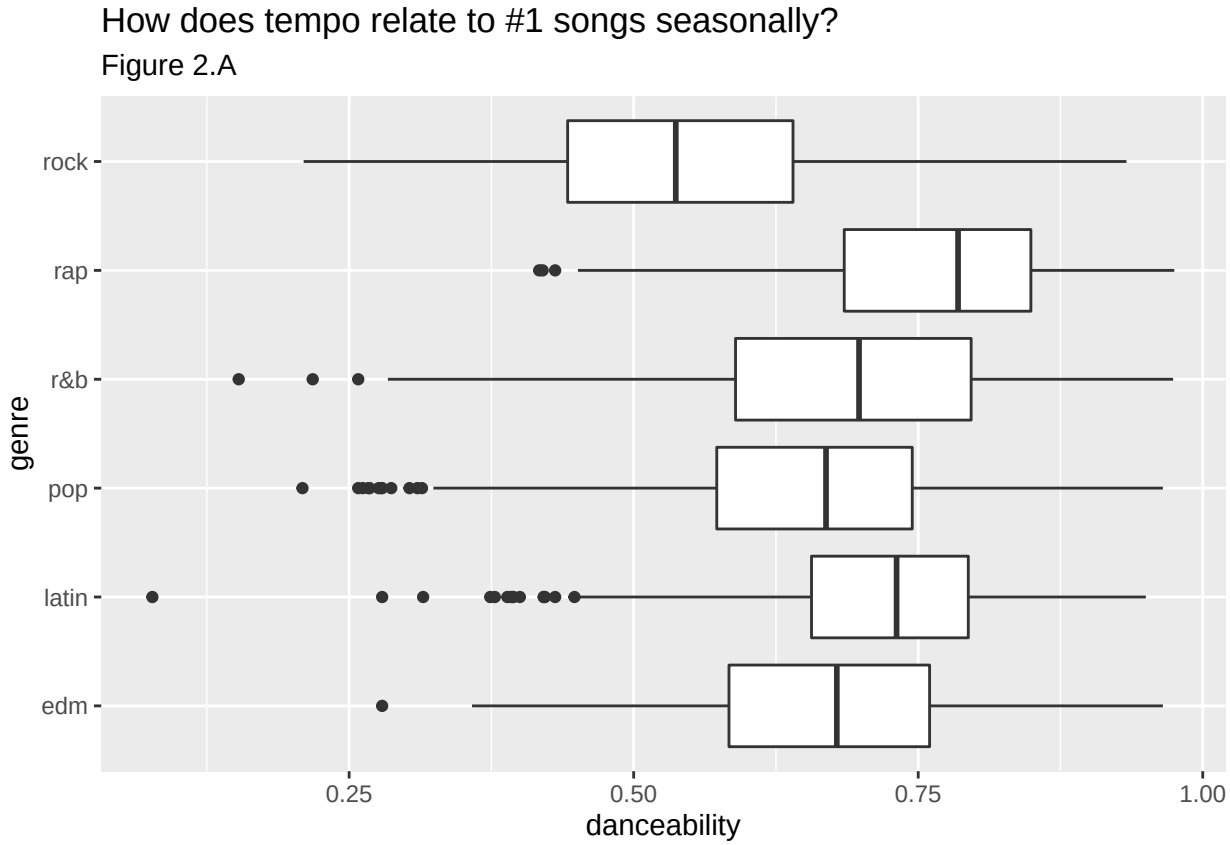
Data

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 a higher value denotes a more popular song). 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

Mark's Stuff:

```
spotify %>% filter(rank == "TOP 10%") %>%
  ggplot(mapping = aes(x = playlist_genre, y = danceability)) +
  geom_boxplot() + coord_flip() +
  labs(title = "How does tempo relate to #1 songs seasonally?",
       subtitle = "Figure 2.A",
       x = "genre", y = "danceability")
```




```
spotify %>% filter(rank == "BOTTOM 90%") %>%
  ggplot(mapping = aes(x = playlist_genre, y = danceability)) +
  geom_boxplot() + coord_flip() +
  labs(title = "How does tempo relate to #1 songs seasonally?",
       subtitle = "Figure 2.A",
       x = "genre", y = "danceability")
```

How does tempo relate to #1 songs seasonally?

Figure 2.A