Spotify Chart Data Analysis

Ezekiel Suarez, Tanu Siddappa

2023-10-13

Introduction

Music is a popular form of entertainment across all groups of people. In the last decade, there has been an increase in the number of genres of music and the overall composition of a "popular" song has changed. The introduction of streaming platforms has also changed how we consume music, the most popular platform being Spotify with over 500 million active users. Streaming platforms have made it easier to share music we like with our friends, create our own personalized playlists, and discover new music or genres we may enjoy.

In this analysis, we will be focusing on Spotify's Top Songs chart, and analyze factors in each song that may contribute to a songs probability to make the charts. The data contains the top 50 songs from each year, from the years 2010-2019. This data set provides information on release year, bpm, danceability, energy, speechiness, and acousticness which can all help in determining factors that may influence a song's popularity. We will be attempting to find correlation between these factors and the charts. We are also interested in how the overall trend of these factors has shifted through the years. Some question we are looking to answer include:

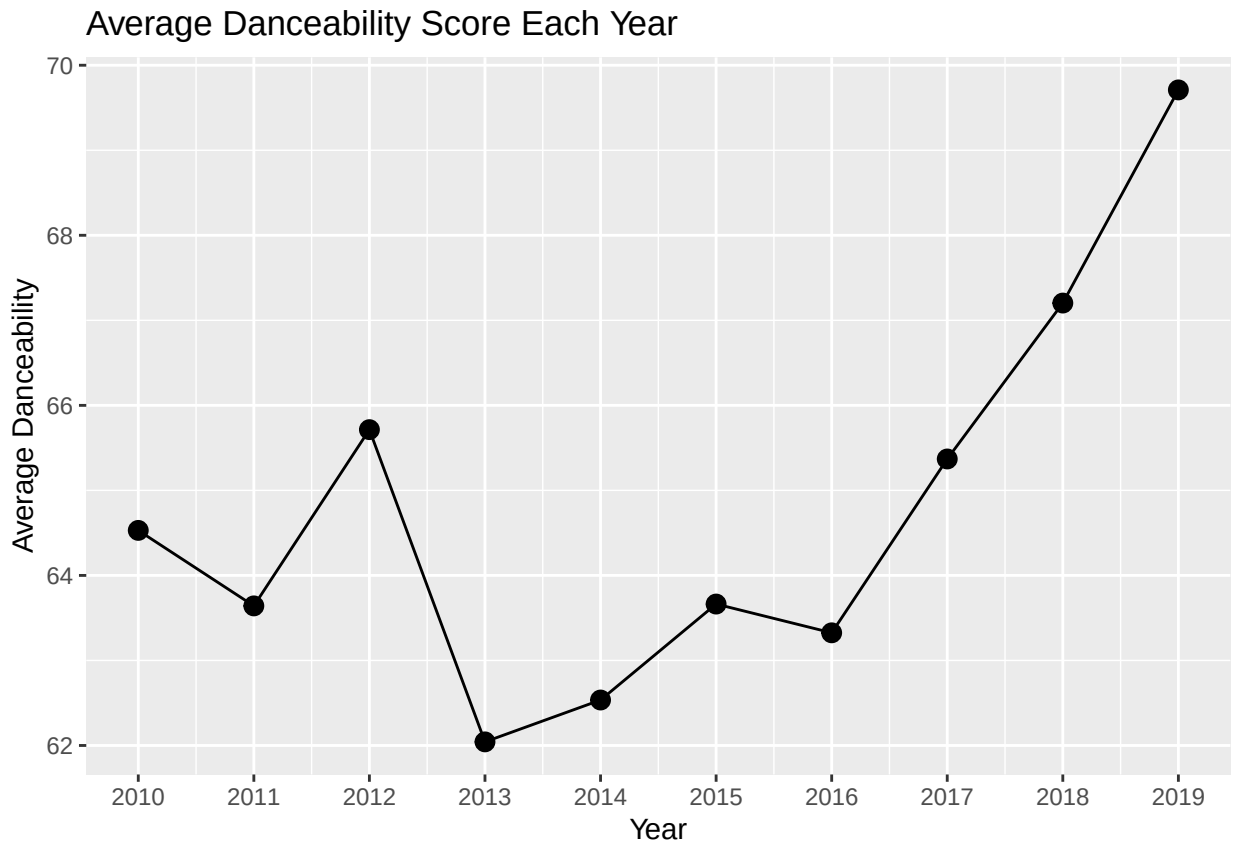
1. How does the length of a song impact a songs popularity?
2. What are the most common genres on the charts?
3. Do more lyrical or more instrumental songs perform better on the Top Songs charts?
4. Which artists are most commonly seen on Top Songs charts?
5. How has the mood of songs on the Top Songs charts shifted through the decade?

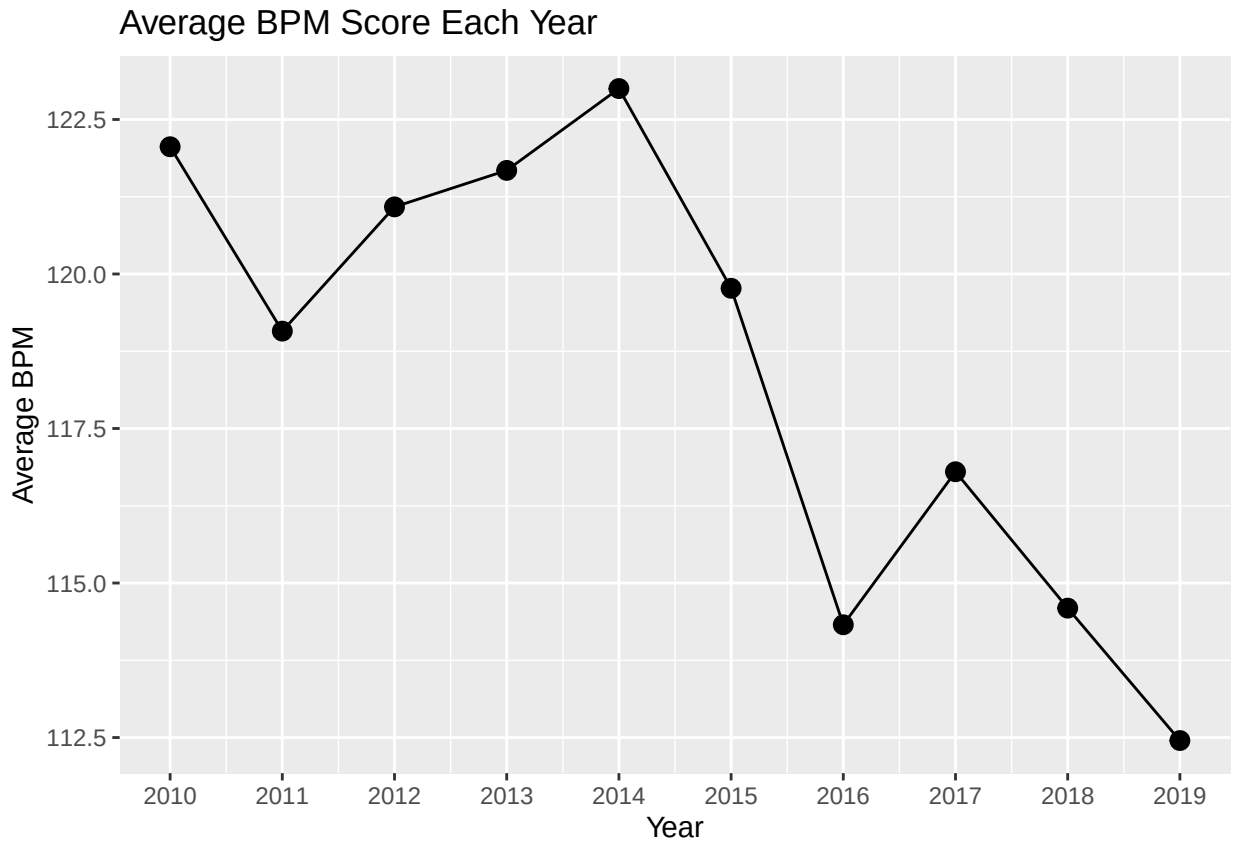
Data Description

There are several variables that the datasets explores about each song. These include:

1. bpm (beats per minute): The tempo of a song
2. nrgy (Energy): The energy level of a song (higher value is more energetic)
3. dnce (Danceablility): How danceable a song is (higher value is more dancable)
4. dB (Loudness): How loud a song is (higher value is more loud)
5. live (Liveness): How likely a song is to be a live recording (higher value is more likely)
6. val (Valence): Positivity of a song (higher value is more positive)
7. dur (Duration): Length of a song in seconds
8. acous (Acousticness): Score for acoustic instrumentation vs digital instrumentation (higher value being more acoustic)
9. spch (Speechiness): Amount of spoken words in a song (higher value has more spoken words)
10. pop (Popularity): Popularity score of a song (higher value is more popular)

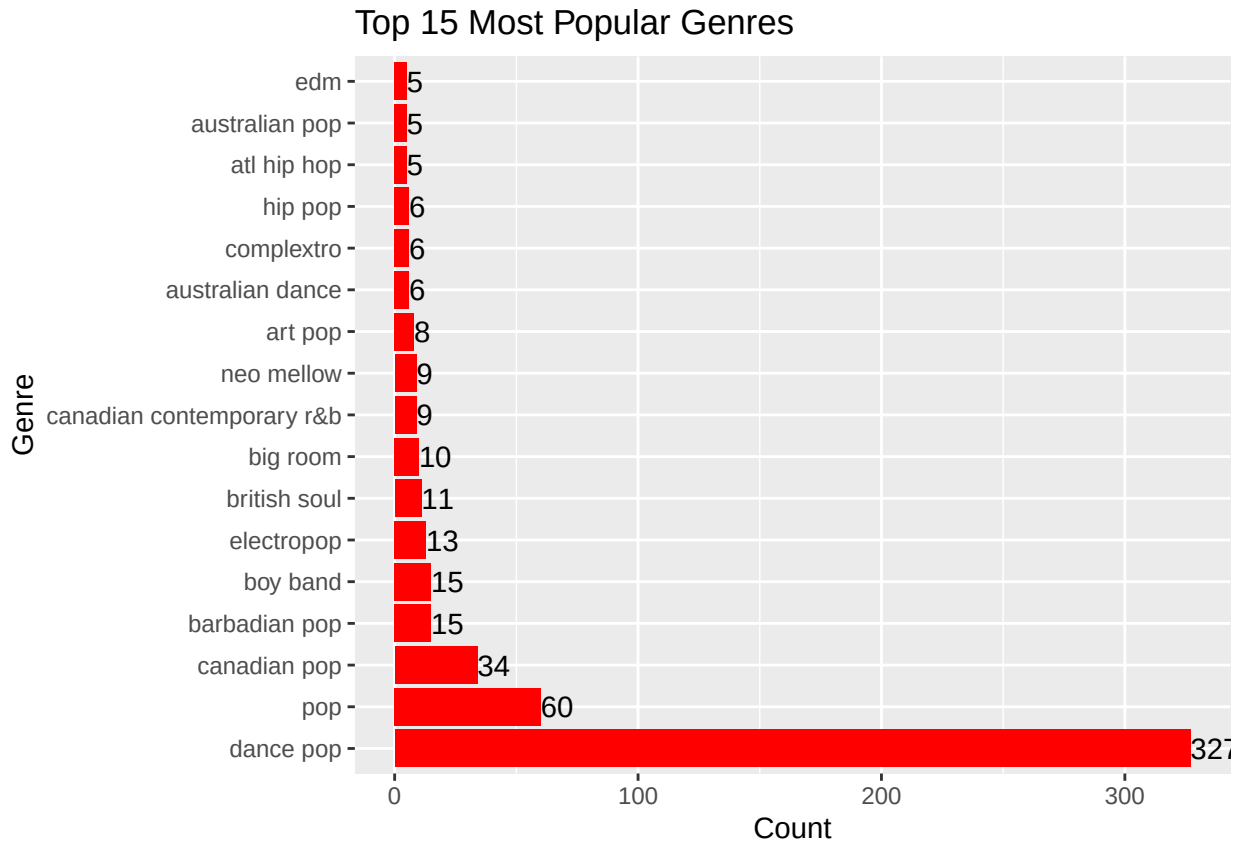
Average Danceability vs Beats Per Minute by Year





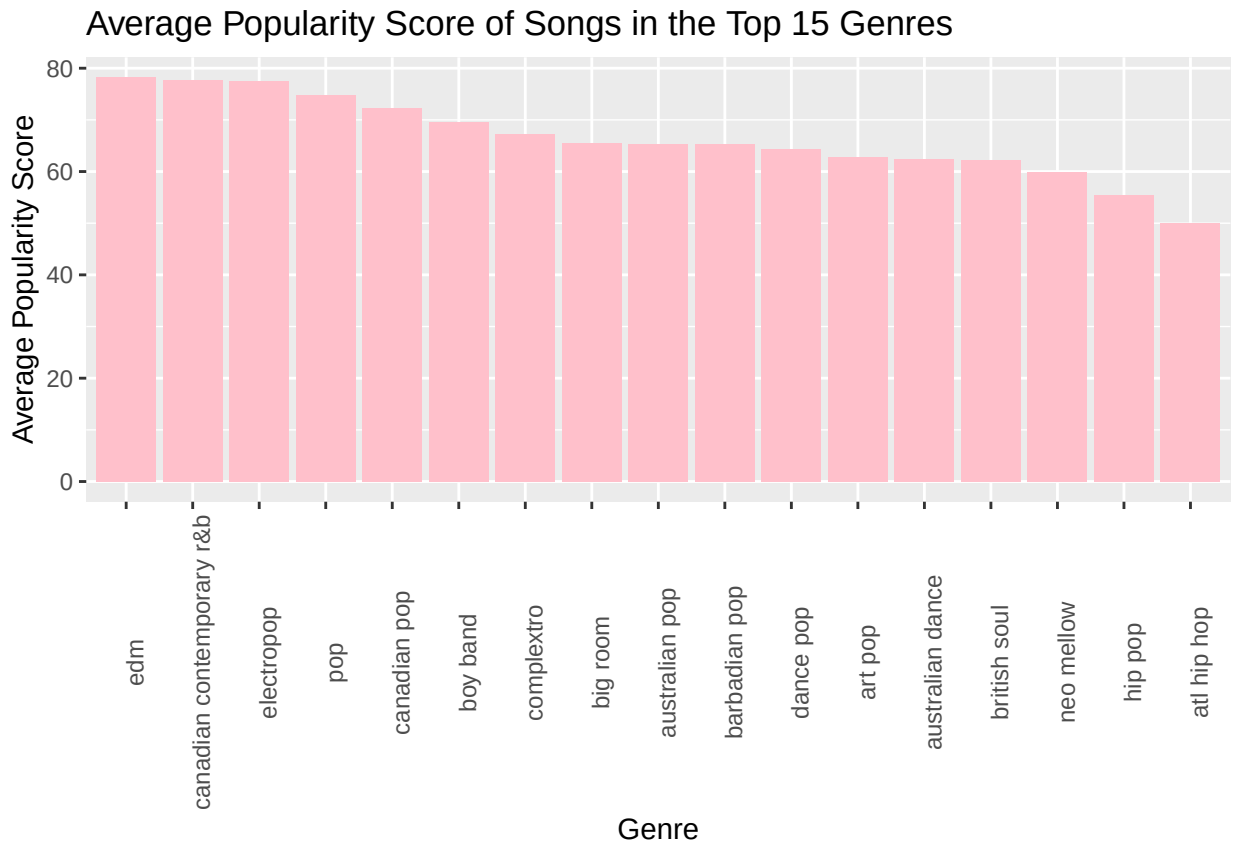
While the danceability of songs on the charts has been increasing since 2016, the average beats per minute has been decreasing overall. This is surprising, because we expect that songs with a faster tempo are more commonly danced to. This shows that over the decade, songs on the charts were more "danceable" even if they had a slower tempo than previous years.

Top 15 Most Popular Genres



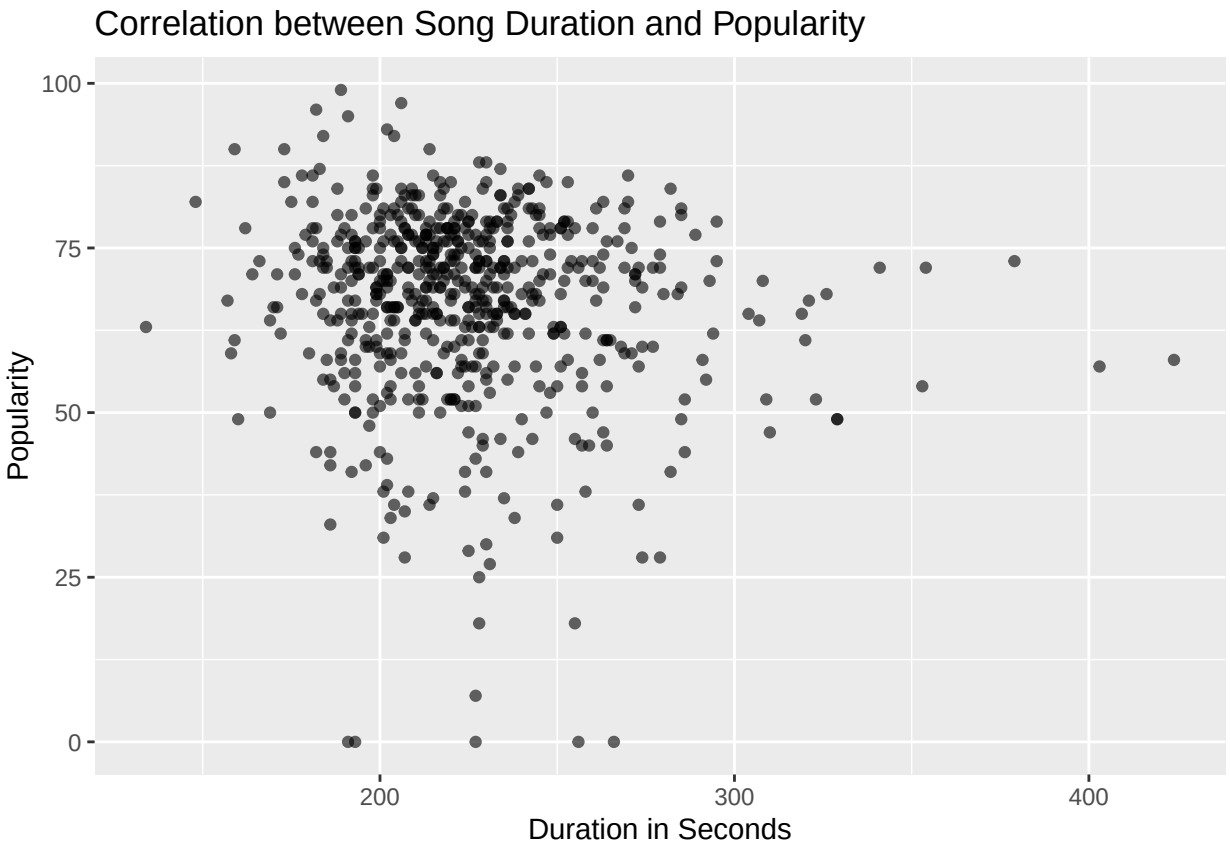
Out of the top 15 most popular genres, the top 4 are all forms of pop music. This is not surprising, since most of the music played on radios is pop music, contributing to it having prevalence on the charts.

Average Popularity Score in the Top 15 Genres



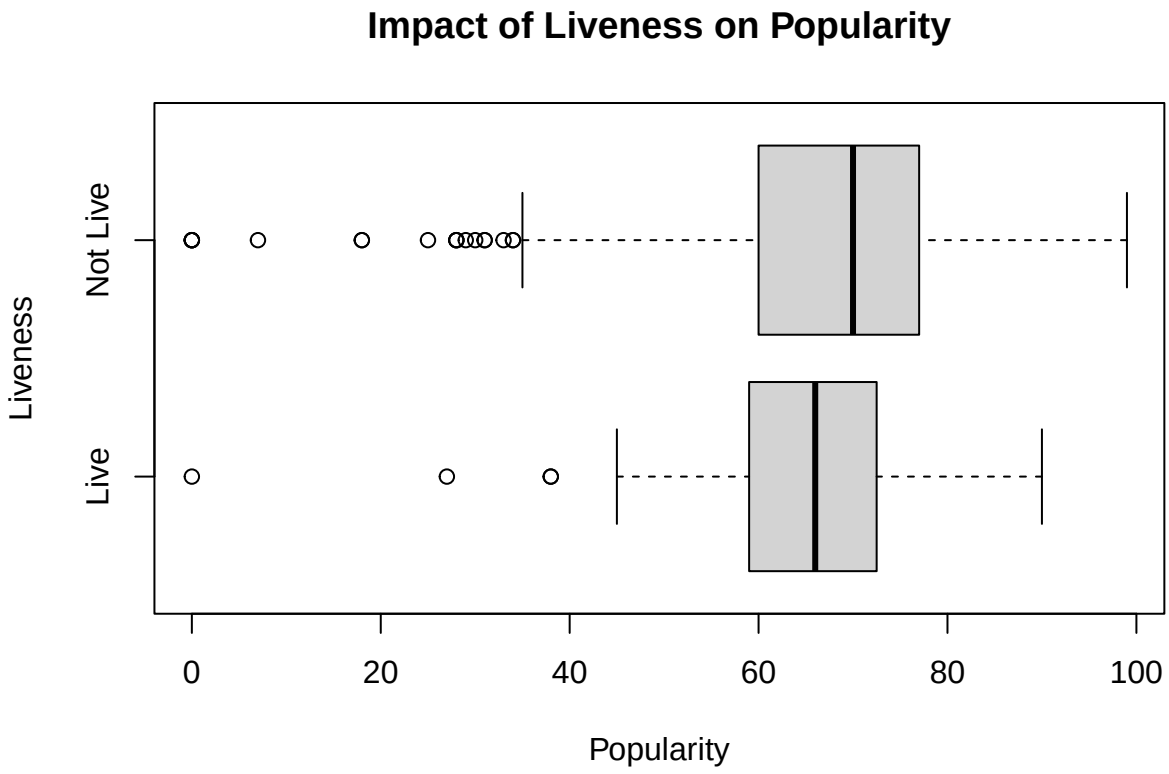
One interesting thing to note in the average popularity score is that popularity is not necessarily corelated to the top genres. Despite having the most dance pop songs on the charts, dance pop is only 11th in popularity score.EDM songs have the highest average popularity score, which could suggest that while less EDM songs make it on the charts overall, the ones that do are very successful.

Correlation between Song Duration and Popularity



There is a cluster of points near the upper-left corner of the plot, which shows that popular songs tend to be shorter in duration (with the exception of a few outliers). This could be because of shorter songs have a better chance to go "viral" on social media. The ideal range of a song to maximize popularity seems to be around 180-210 seconds.

Popularity of songs based on Liveness

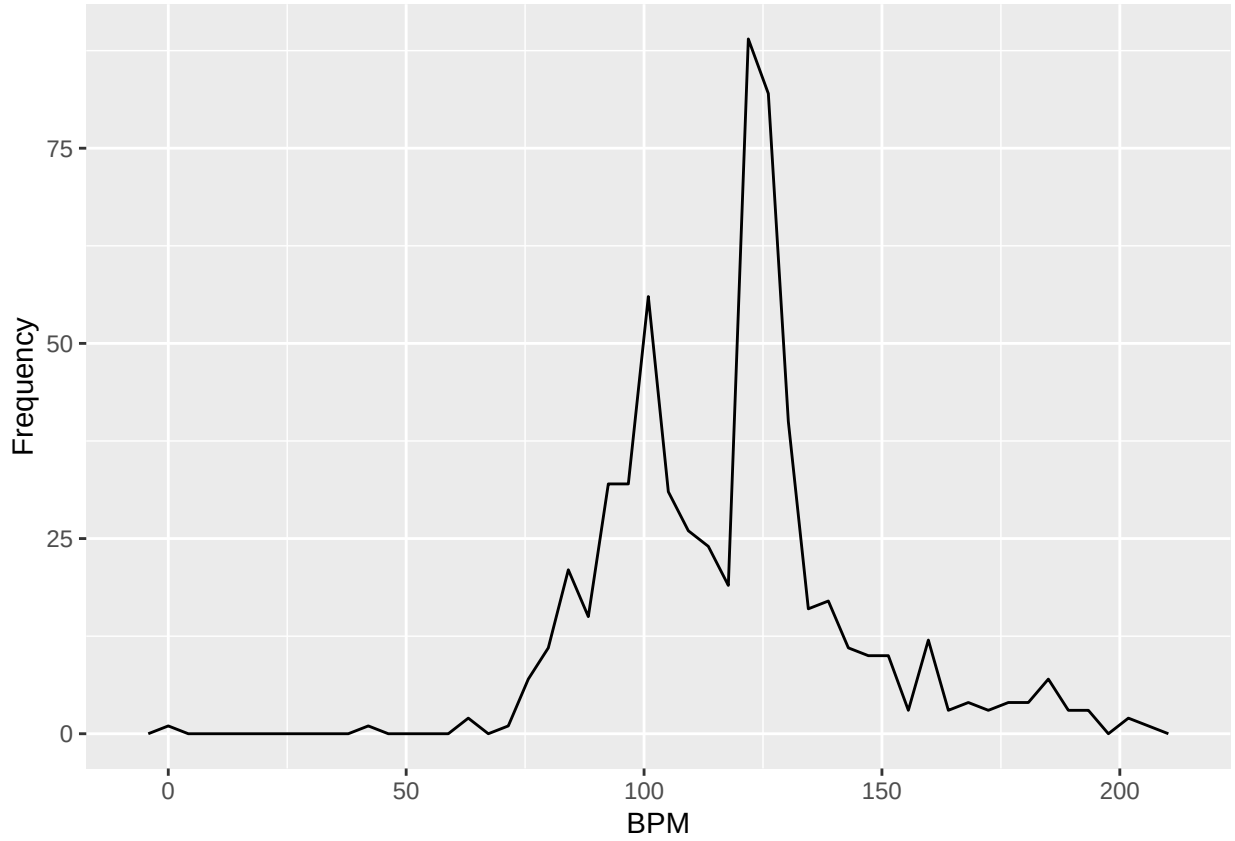


For this plot, songs with a live score of 37 or more are considered "live" songs. Songs that are considerd live tend to be less average than non-live songs. Regularly recorded songs also have more spread in their popularity, and have a much larger maximum popularity score.

Top 15 Artists on Charts

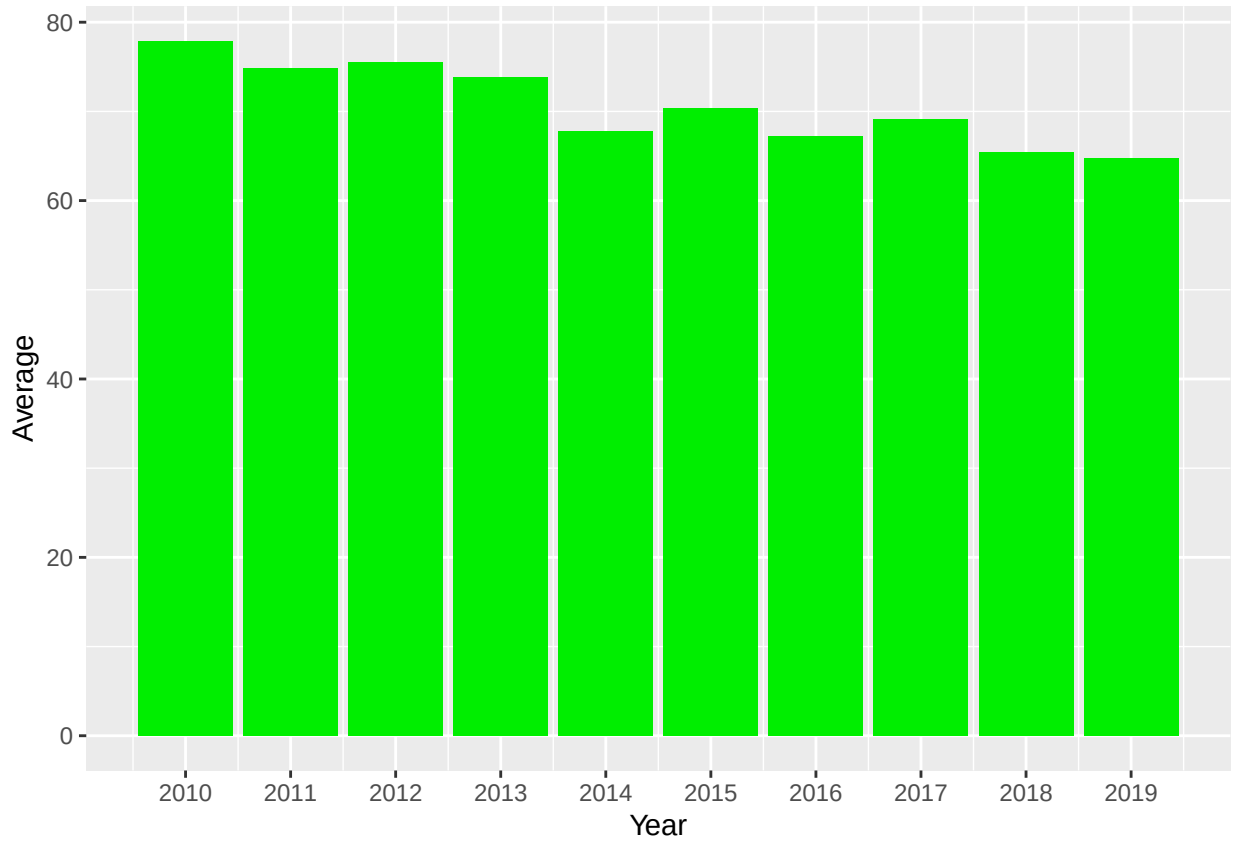
```
##
##       Katy Perry    Justin Bieber         Maroon 5          Rihanna
##               17               16               15               15
##        Lady Gaga       Bruno Mars       Ed Sheeran          Pitbull
##               14               13               11               11
##     Shawn Mendes The Chainsmokers            Adele    Calvin Harris
##               11               11               10               10
##   Jennifer Lopez    Ariana Grande   Britney Spears
##               10                9                9
```

Significance of BPM in Popular Music



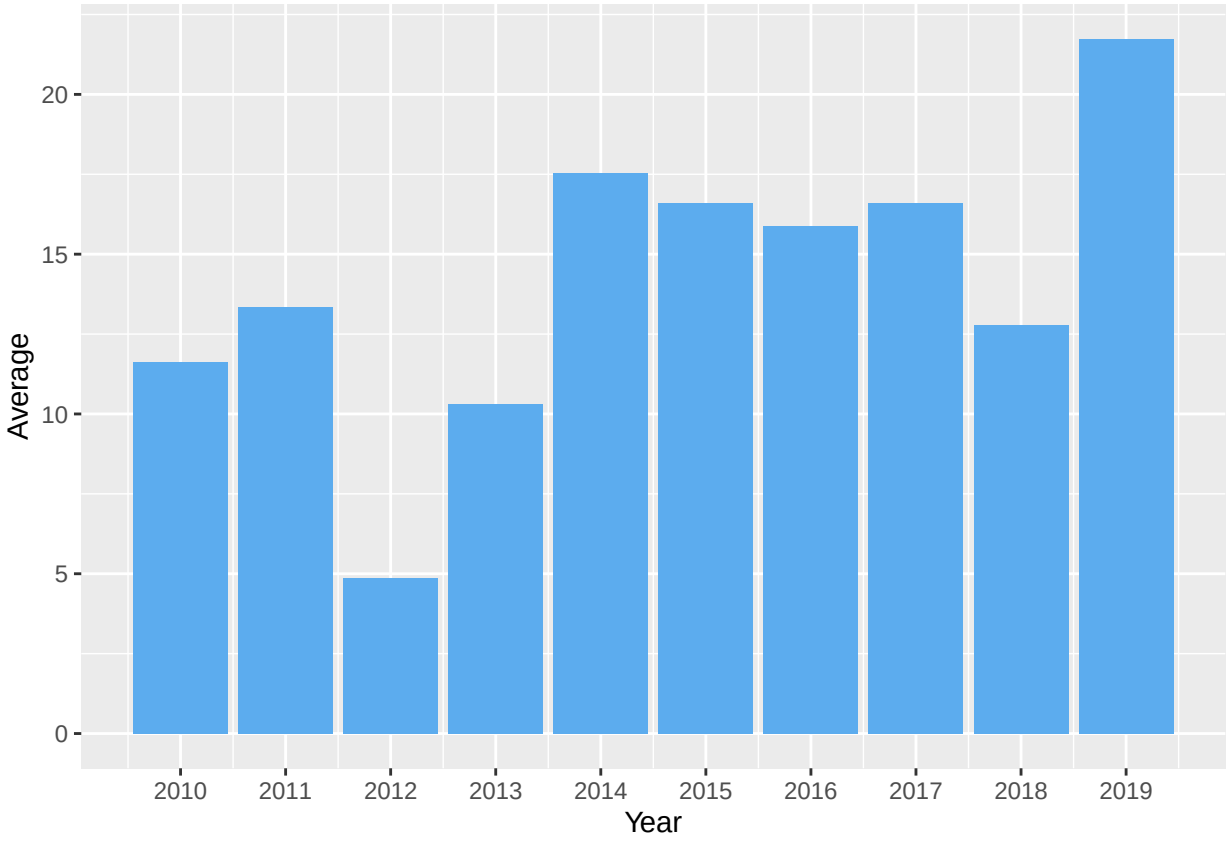
Based on the beats per minute of the previous decade's charting songs, 100 or 120-130 BPM seems to be a solid tempo. This would make for a medium-high pace which makes sense when thinking about the pop genre and what it should sound like.

Song Energy Over Time



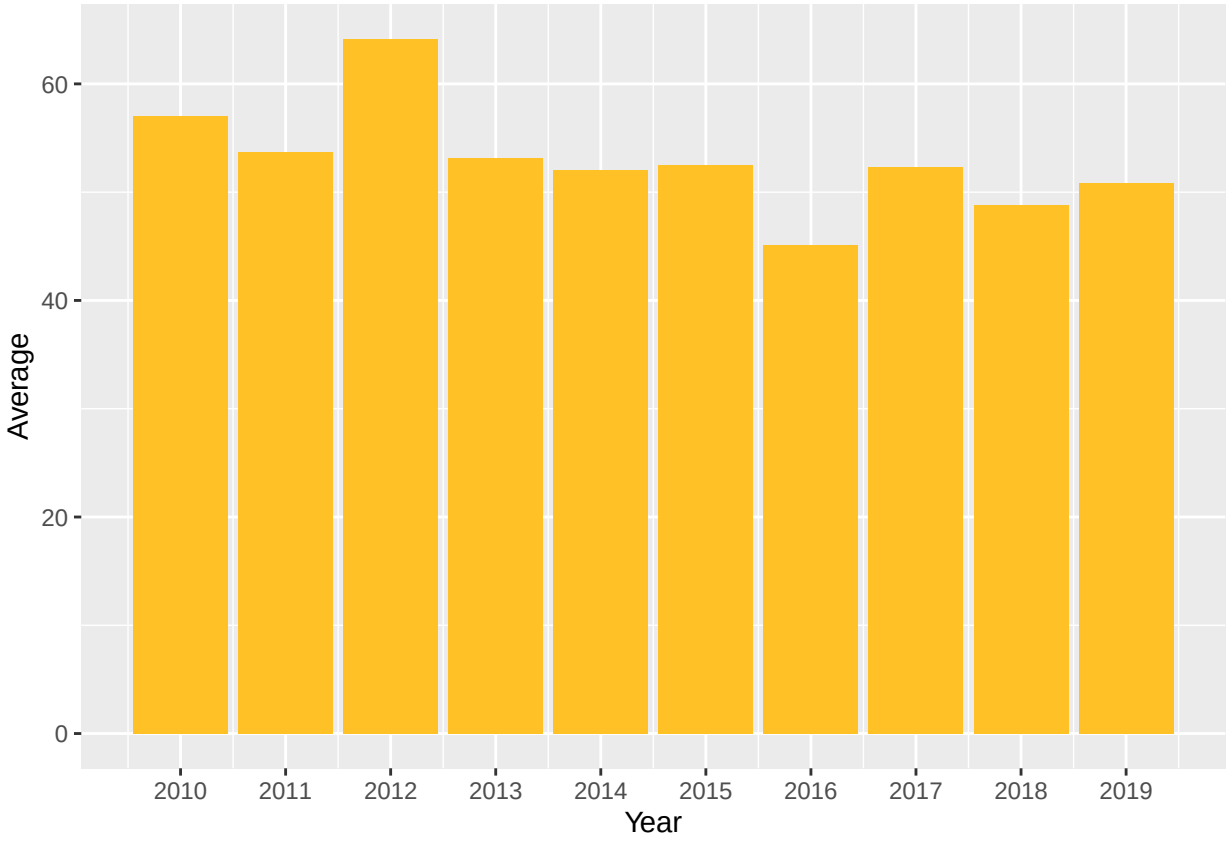
Between 2010 and 2020, song "energy" has decreased which probably means culture as a whole may have gotten more depressive. Sad musicians such as Mitski, Lana Del Ray, and Billie Eilish seem to be in the perfect spot right now.

Acousticness Over Time



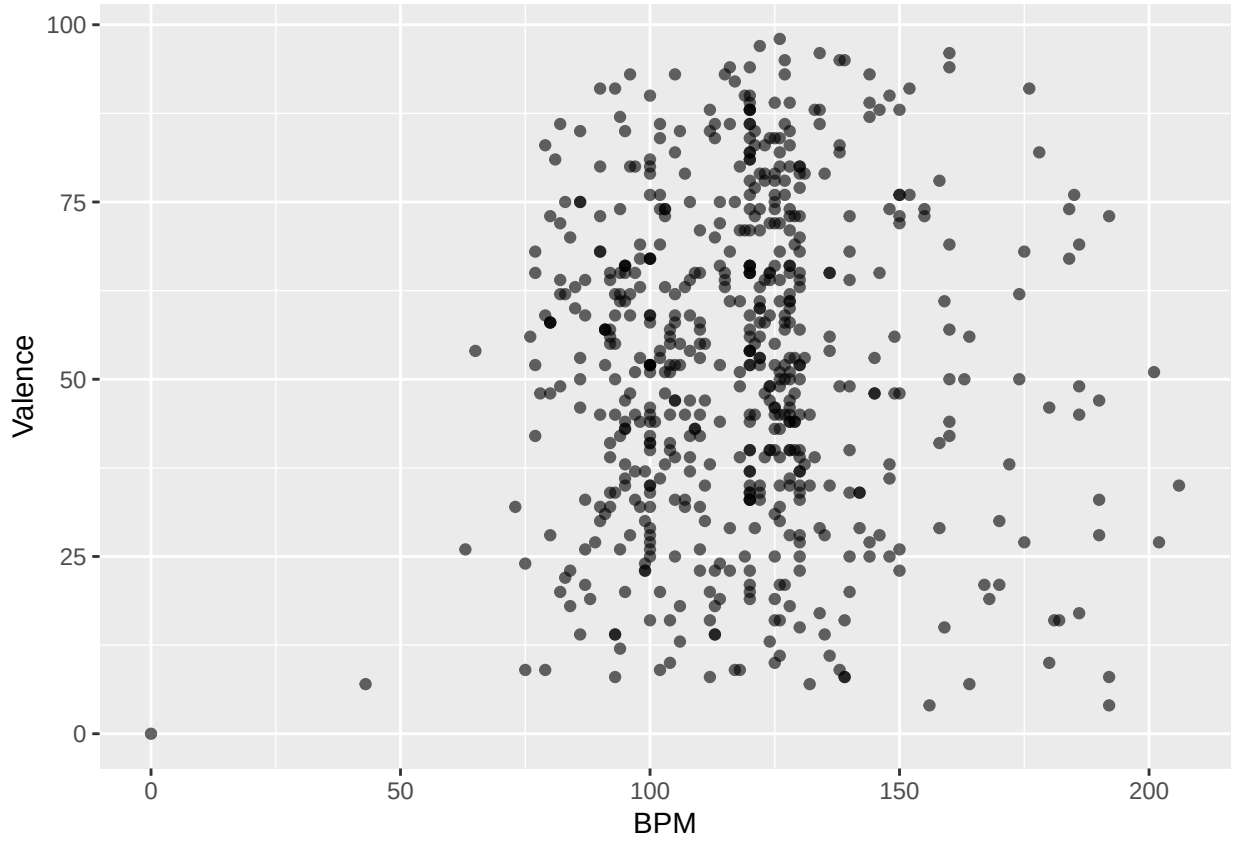
Additionally acoustics have risen dramatically, which can be linked to a quickly changing cultural landscape. These sadder musicians sound more endearing on an acoustic guitar than a MIDI drum loop. Additionally, I think the dramatic change from 2012 to 2014 has to do with the folk scene where bands like The Postal Service, Of Monsters and Men, and Hozier were starting to make an impact.

Valence Over Time



Emotional content has gotten slightly more depressive as guessed, but the happiness of the most in-your-face songs like HAPPY by Pharrell Williams or Can't Stop the Feeling by Justin Timberlake seem to boost scores here.

Energy vs Valence



Surprisingly, beats per minute and happiness of songs have little-to-no correlation and it cannot be said that to have a happy song, it has to be fast or to have a sad song, it has to be slow.

Includes Featured Artists

```
## [1] "13.4328358208955 %"
```

Under 15% of songs overall have featured artists meaning unless the featured is more popular than the main artist, features are not very important to chartability. HOWEVER, dance pop usually will have a feature due to the nature of dance music. DJ Khaled, Marshmallow, or Skrillex will not be singing, so they need to hire people to sing over their beats.

Dance Pop[ularity]

```
## [1] "54.228855721393 %"
```

Dance pop takes up a whopping 54% of the 2010s music charts, showing the impact of following a trend. If people want a certain sound, they are going to listen to it even if it has already been done 327 times. Rap and hip hop have a surprisingly low amount of songs, even though it is definitely heard on the radio. This means that rapping has become more of a norm in pop music and is less of the independent genre it used to be. “hip pop”, “atl hip hop”, and “hip hop” only made it to positions 14, 15, and 18 respectively.

```
##       top.genre   n
## 1     dance pop 327
## 2           pop  60
## 3  canadian pop  34
```

```
## 4 barbadian pop  15
## 5      boy band  15
```

Conclusion

Music is everchanging and people often are as well. A popular song is often easy to listen to, and will have some amount of familiarity mixed with a bit of fresh new noise. We can gather from the last decade that year by year, not much drastically will change, but depending on current events, certain moods or tempos may be more gravitated toward.

In the last decade, there was a decrease in the average BPM of songs on the charts, meaning that more slow-paced songs were on the charts. The most popular songs also tended to be shorter in duration. Pop was prevalent on the charts throughout the decade, and was the most popular genre on the charts by far. Live songs also tended to have less spread in their popularity than non-live songs. This could be because of the studio finish of live songs, and their radio presence. From all of these plots, we can predict that shorter pop songs that are medium paced will be successful on Spotify's charts.

Appendix

Top Spotify Songs Dataset: https://www.kaggle.com/datasets/leonardopena/top-spotify-songs-from-20102019-by-year/data

The plots in this project are all generated from the following code:

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(ggplot2)
library(plotly)
library(dplyr)
library(knitr)
library(stringr)
Spotify <- read.csv("top10s.csv")
##Spotify = read.csv("top10s.csv",sep=";",header=TRUE)

avgDance<- Spotify %>%
  group_by(year) %>%
  summarize(average = mean(dnce, na.rm = TRUE))

ggplot(data = avgDance, aes(x = year, y = average)) +
  geom_point(size=3) +
  geom_line()+
  labs(title = "Average Danceability Score Each Year",
       x = "Year",
       y = "Average Danceability") +
  scale_x_continuous(breaks = seq(2010, 2019, by = 1))

avgBpm<- Spotify %>%
  group_by(year) %>%
  summarize(average = mean(bpm, na.rm = TRUE))

 ggplot(data = avgBpm, aes(x = year, y = average)) +
  geom_point(size=3) +
  geom_line()+
  labs(title = "Average BPM Score Each Year",
       x = "Year",
       y = "Average BPM") +
```

```
  scale_x_continuous(breaks = seq(2010, 2019, by = 1))

genres <- Spotify %>%
  group_by(top.genre) %>%
  summarize(nums = n()) %>%
  slice_max(order_by = nums, n = 15)

  ggplot(data = genres, aes(x = reorder(top.genre, -nums), y = nums)) +
  coord_flip() +
  geom_text(aes(label = nums), hjust = 0) +
  geom_bar(stat = "identity", fill="red") +
  labs(title = "Top 15 Most Popular Genres", x = "Genre", y = "Count")


genre <- Spotify %>%
  group_by(`top.genre`) %>%
  summarize(nums = n()) %>%
  slice_max(order_by = nums, n = 15) %>%
  pull(`top.genre`)

averagePop <- Spotify %>%
  filter(`top.genre` %in% genre) %>%
  group_by(`top.genre`) %>%
  summarize(average = mean(pop, na.rm = TRUE))

ggplot(averagePop, aes(x = reorder(`top.genre`, -average), y = average)) +
  geom_bar(stat = "identity", fill="pink") +
  labs(title = "Average Popularity Score of Songs in the Top 15 Genres", x = "Genre",
       y = "Average Popularity Score") +
  theme(axis.text.x = element_text(angle = 90))

ggplot(data = Spotify, aes(x = dur, y = pop)) +
  geom_point(alpha = 0.6) +
  labs(
    title = "Correlation between Song Duration and Popularity",
    x = "Duration in Seconds",
    y = "Popularity")
Spotify$liveBinary = ifelse(Spotify$live > 37, "Live", "Not Live")
boxplot(pop ~ liveBinary, data = Spotify,
        main="Impact of Liveness on Popularity",
        xlab="Popularity",
        ylab="Liveness",
        horizontal=TRUE)
nums = table(Spotify$artist)
sorted = sort(nums, decreasing = TRUE)
head(sorted,15)
ggplot(Spotify, aes(x=bpm)) + geom_freqpoly(bins=50)+labs(x="BPM", y="Frequency")

nrgyVsYear = Spotify %>%
  group_by(year) %>%
  summarise(Average = mean(nrgy))

ggplot(nrgyVsYear, aes(x=year,y=Average)) + geom_bar(stat = "identity", fill="green2") +
```

```
  labs(x="Year") + scale_x_continuous(breaks = seq(2010, 2019, by = 1))

acousVsYear = Spotify %>%
  group_by(year) %>%
  summarise(Average = mean(acous))

ggplot(acousVsYear, aes(x=year,y=Average)) + geom_bar(stat = "identity", fill="steelblue2") +
  labs(x="Year") + scale_x_continuous(breaks = seq(2010, 2019, by = 1))
valVsYear = Spotify %>%
  group_by(year) %>%
  summarise(Average = mean(val))

ggplot(valVsYear, aes(x=year,y=Average)) + geom_bar(stat = "identity", fill="goldenrod1") +
  labs(x="Year") + scale_x_continuous(breaks = seq(2010, 2019, by = 1))
ggplot(data = Spotify, aes(x = bpm, y = val)) +
  geom_point(alpha = 0.6) +
  labs(
    x = "BPM",
    y = "Valence")

feats = count(Spotify %>% filter(str_detect(title, "feat")))
songs = nrow(Spotify)
div = feats/songs
print(paste(div*100, "%"))
feats = count(Spotify %>% filter(str_detect(`top.genre`, "dance pop")))
songs = nrow(Spotify)
div = feats/songs
print(paste(div*100, "%"))
Spotify %>% count(`top.genre`) %>% arrange(desc(n)) %>% head(5)
```