Project proposal

Approximately Normal: Jenny Huang, Bella Larsen, Jessie Bierschenk, Aidan Gildea

6 October 2020

```
library(tidyverse)
library(broom)
library(patchwork)
```

```
spotify <- read_csv("data/spotify_songs.csv")
```

Section 1. Introduction

With streaming services dominating the music industry in today's world, Spotify has become a popular platform for individuals to listen to an unlimited variety of songs. In our study, we are investigating different characteristics of about 32,000 songs on Spotify. The variables of concern pertain to logistical aspects of the song (release date, playlist genre, etc.) as well as technical qualities of the song itself (tempo, rhythm, energy, loudness, etc.). It proves difficult to determine or forecast a given song's popularity through mere anecdotal experience or personal opinion; by investigating both qualitative and quantitative aspects of Spotify songs, we hope to gain a more precise understanding of what the common characteristics are that render a song popular. Motivation for our research question draws from how the nuances of particular songs can affect their popularity, as well as how audio features are perceived as appealing and what makes a song popular outside of its genre. In the article "Understanding + classifying genres using Spotify audio features," written by data scientist Kayla Pavlik, quantifiable ratings are discussed that differentiate audible components of a song [1]. We hope to include quantifiable characteristics of audible features to see if there are notable trends of more popular songs, or if traits such as genre and artists prove to be a more reliable indicator of a song's popularity as suggested by the article "Visualized: Can We Quantify the Most Popular Music?" by Madeleine Picard [2].

We are interested in finding out about which audio features of a song on Spotify predict its popularity. We will be using the `track_popularity` as the response variable, a numerical score that ranges 0 to 100, with 100 being the best score. We will use numerical predictor variables and categorical predictor variables to predict track popularity [3]. Of the categorical variables, we believe that `playlist_genre` will be a top predictor of `track_popularity`, as rock, pop, and country songs are highly popular among listeners [4]. For our numerical variables, we hypothesize that songs that score high in danceability and energy tend to be more popular based on our experiences with popular music.

Section 2. Data description

```
observations <- as.tibble(count(spotify)) %>%
  mutate(variables = ncol(spotify))
observations %>%
  kable()
```

n	variables
32833	23

In our data set `spotify`, there are 32,833 observations, each of which is a song/track on Spotify. There are

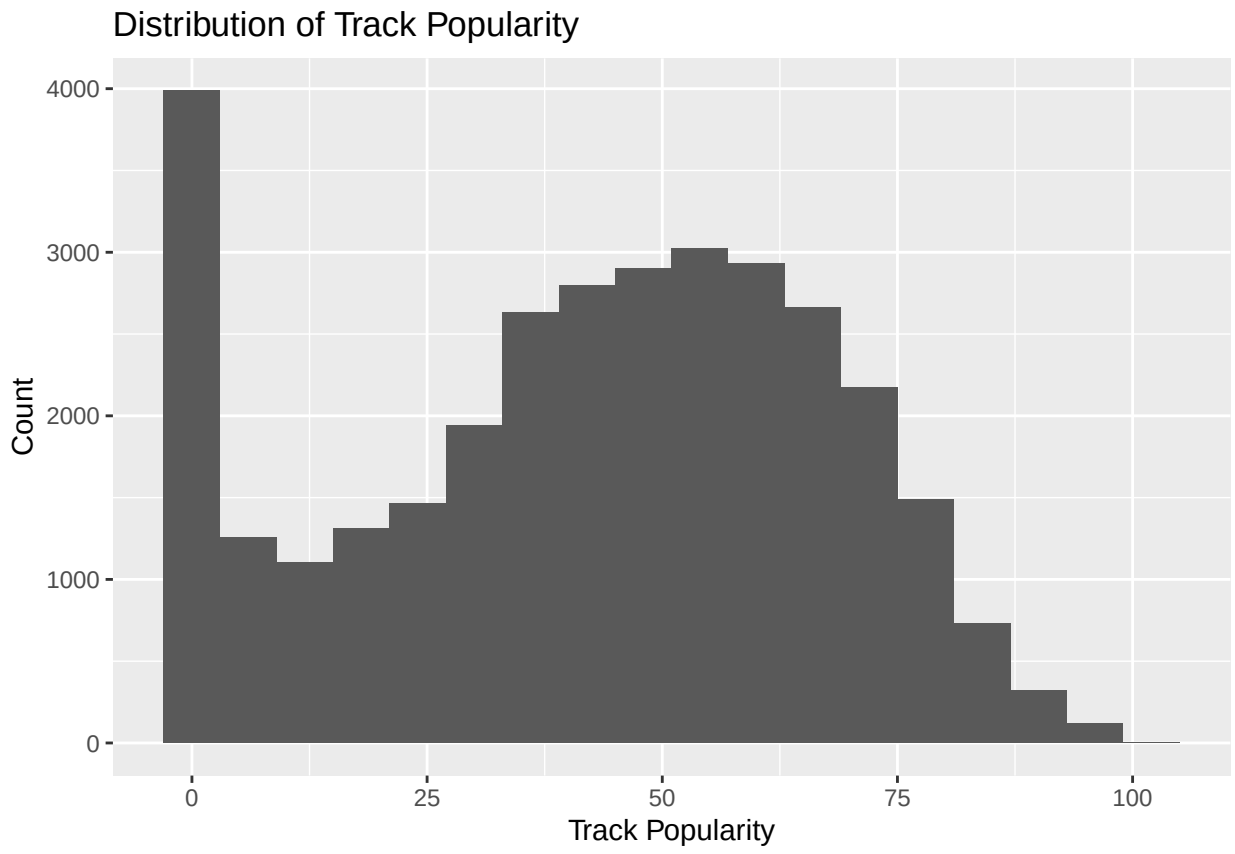
23 variables in the data set, ranging from the song's release date to its danceability.

```
spotify[,c("track_popularity", "playlist_genre")] %>%
  arrange(desc(track_popularity)) %>%
  slice(1:10) %>%
  kable()
```

track_popularity	playlist_genre
100	pop
100	latin
99	latin
99	r&b
99	r&b
99	edm
98	pop
98	pop
98	pop
98	pop

The primary response variable of our regression analysis will be `track_popularity`. The variable is numeric and ranges from values of 0 (least popular) to 100 (most popular). The popularity is reportedly measured by the amount of streams for a given track [5]. A distribution of `track_popularity` can be seen below.

```
ggplot(data=spotify, aes(x=track_popularity)) +
  geom_histogram(binwidth = 6) +
  labs(title= "Distribution of Track Popularity", x= "Track Popularity", y = "Count")
```



```
spotify %>%
  summarise(min = min(track_popularity),
            q1 = quantile(track_popularity, probs = .25),
            median = median(track_popularity),
            q3 = quantile(track_popularity, probs = .75),
            max = max(track_popularity),
            iqr = q3-q1,
            mean = mean(track_popularity),
            std_dev = sd(track_popularity)
            ) %>%
  kable()
```

min	q1	median	q3	max	iqr	mean	std_dev
0	24	45	62	100	38	42.47708	24.98407

```
spotify %>%
count(track_popularity == 0)
```

```
## # A tibble: 2 x 2
##   `track_popularity == 0`     n
##   <lgl>                   <int>
## 1 FALSE                   30130
## 2 TRUE                     2703
```

The distribution of `track_popularity` is bimodal and centered at a mean of 42.48. The distribution has a standard deviation of 24.98, which measures its spread. There are 2703 songs that have a `track_popularity` of 0.

To predict the popularity of a song, we will consider the variables that describe the energy, loudness (measured in decibels), danceability, speechiness, duration (measured in milliseconds), mode, acousticness, valence, tempo (beats per minute), and playlist genre of each track observation. Energy, danceability, speechiness, acousticness, and valence are all measured on a scale from 0 to 1. These predictor variables are sourced from Spotify's application programming interface (API) available online [6]. Energy is a variable that describes the intensity of a track, where 0 describes low-energy songs and 1 describes high-energy songs [6]. Danceability is defined as "how suitable a track is for dancing based on a combination of musical elements," where 0 describes songs that are least danceable and 1 describes songs that are most danceable [6]. Speechiness describes the amount of words in a track, where tracks that are fully instrumental would have a value closer to 0 and tracks that are fully spoken word (such as a podcast) would have a value closer to 1 [6]. Mode is a measurement of the major or minor key in a track, where 1 is major and 0 is minor [6]. Acousticness measures confidence of whether a track is acoustic, where a track with a value of 1 would have "high confidence" of being acoustic (not electronic), and a song with a value of 0 would have low confidence of being acoustic [6]. Valence is a measure of the positivity of a track, where 0 describes songs that are low-valence (sad, angry), while 1 describes more positive or happy songs [6].

These data were collected by Spotify's developers and data scientists as part of Spotify's API. The dataset consists of audio features for different tracks on Spotify that have been measured by Spotify analytics for each available track and made publicly available on the Spotify API [7]. The latest song included in this dataset was released January 29th, 2020 [3]. Kaylin Pavlik published an analysis of these data using the "spotifyr" package to classify genre of tracks as part of the Tidy Tuesday project that made these data available on GitHub [3,8].

Section 3. Glimpse of data

```
glimpse(spotify)
```

```
## Rows: 32,833
## Columns: 23
## $ track_id                <chr> "6f807x0ima9a1j3VPbc7VN", "0r7CVbZTWZgbTCY...
## $ track_name              <chr> "I Don't Care (with Justin Bieber) - Loud ...
## $ track_artist            <chr> "Ed Sheeran", "Maroon 5", "Zara Larsson", ...
## $ track_popularity        <dbl> 66, 67, 70, 60, 69, 67, 62, 69, 68, 67, 58...
## $ track_album_id          <chr> "2oCs0DGTsRO98Gh5ZSl2Cx", "63rPSO264uRjW1X...
## $ track_album_name        <chr> "I Don't Care (with Justin Bieber) [Loud L...
## $ track_album_release_date <chr> "2019-06-14", "2019-12-13", "2019-07-05", ...
## $ playlist_name           <chr> "Pop Remix", "Pop Remix", "Pop Remix", "Po...
## $ playlist_id             <chr> "37i9dQZF1DXcZDD7cfEKhW", "37i9dQZF1DXcZDD...
## $ playlist_genre          <chr> "pop", "pop", "pop", "pop", "pop", "pop", ...
## $ playlist_subgenre       <chr> "dance pop", "dance pop", "dance pop", "da...
## $ danceability            <dbl> 0.748, 0.726, 0.675, 0.718, 0.650, 0.675, ...
## $ energy                  <dbl> 0.916, 0.815, 0.931, 0.930, 0.833, 0.919, ...
## $ key                     <dbl> 6, 11, 1, 7, 1, 8, 5, 4, 8, 2, 6, 8, 1, 5,...
## $ loudness                <dbl> -2.634, -4.969, -3.432, -3.778, -4.672, -5...
## $ mode                    <dbl> 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, ...
## $ speechiness             <dbl> 0.0583, 0.0373, 0.0742, 0.1020, 0.0359, 0....
## $ acousticness            <dbl> 0.10200, 0.07240, 0.07940, 0.02870, 0.0803...
## $ instrumentalness        <dbl> 0.00e+00, 4.21e-03, 2.33e-05, 9.43e-06, 0....
## $ liveness                <dbl> 0.0653, 0.3570, 0.1100, 0.2040, 0.0833, 0....
## $ valence                 <dbl> 0.518, 0.693, 0.613, 0.277, 0.725, 0.585, ...
## $ tempo                   <dbl> 122.036, 99.972, 124.008, 121.956, 123.976...
## $ duration_ms             <dbl> 194754, 162600, 176616, 169093, 189052, 16...
```

References

[1] https://www.kaylinpavlik.com/classifying-songs-genres/ [2] https://www.displayr.com/most-popular-music/ [3] https://github.com/rfordatascience/tidytuesday/tree/master/data/2020/2020-01-21 [4] https://www.statista.com/statistics/442354/music-genres-preferred-consumers-usa/ [5] https://stackoverflow.com/questions/17727208/spotify-how-is-a-tracks-popularity-value-determined [6] https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/ [7] https://developer.spotify.com/documentation/web-api/ [8] https://www.rdocumentation.org/packages/spotifyr/versions/1.0.0