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

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

```
spotify[,c("playlist_genre", "playlist_subgenre")]
```

```
## # A tibble: 32,833 x 2
##    playlist_genre playlist_subgenre
##    <chr>          <chr>
##  1 pop            dance pop
##  2 pop            dance pop
##  3 pop            dance pop
##  4 pop            dance pop
##  5 pop            dance pop
##  6 pop            dance pop
##  7 pop            dance pop
##  8 pop            dance pop
##  9 pop            dance pop
## 10 pop            dance pop
## # ... with 32,823 more rows
```

The primary response variable of our regression analysis will be `playlist_genre`. Using the other characteristic variables of the songs ('key', 'loudness', 'acousticness, etc.), we will predict the genre of the playlist each song would be placed in. Depending on the criteria of the project, we may also extend our analysis to predict the subgenre of the playlist a song is in based off its characteristics.

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...
```