Project proposal

Fantastic FouR: Hannah, Eli, Preetha

2020-10-07

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

```
spotify_songs <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/
```

Section 1. Introduction

In this project, we will be looking at music on Spotify. While every individual has unique tastes in music, there are still songs which are much more popular than others. We each independently noticed that the music that goes number 1 on the charts sounds similar even if two different artists make it. Thus, we wanted to explore if there is a formula for making popular music, or if our conceptions are exaggerated. We will do this by investigating various characteristics of sound for many songs on Spotify.
The general research question we wish to explore is as follows: what characteristics best predict the popularity of a song?
We hypothesize that danceability, energy, and loudness are the strongest predictors of the popularity of a song.

Section 2. Data description

There are 33,000+ songs in this dataset. Each observation represents a unique song scored on scales of danceability, energy, loudness, mods, speechiness, acousticness, instrumentalness, liveness, valence, tempo, and duration_ms. These scales range from values of 0 to 1. Each observation also has various descriptors associated with it, including track popularity, key, album name, album release data, and information associated with playlists the song is on. The variable that we have designated as the response is track_popularity, which is a score from 0 to 100 measuring the song's overall popularity. The raw data was collected directly from Spotify, which uses machine learning algorithms to evaluate key parameters for each song, such as danceability, energy, etc. Using the package spotifyr, these data points were pulled and saved in a csv file to be used for analysis.

There are several predictor variables in this dataset, including descriptors (like track and album metadata as well as names for the song and playlists, etc.), continuous variable scores on scales of 0-1.0, and categorical variables regarding modality and key. They are as follows: track_name refers to the name of the song. We may use this to identify the most popular songs at the end of our analysis. Track_artist refers, of course, to the song artist. We may use this, like track name, to identify if a few artists release the most popular songs (in our dataset). Track_album_name refers to the album from which a particular song is from. We may use this to identify if many of the most popular songs come from the same album. Track_album_release_date refers to the date when an album is released. We may use this data to filter for songs that are more recent, to cut down on the 33,000+ observations that are currently in the dataset to something more manageable.
Danceability refers to how suitable a track is for dancing, a metric that combines assessments of tempo, beat strength, regularity, etc. This is scored on a scale from 0.0-1.0, with 0.0 being the least danceable.
Energy is a measure of how energetic (fast, loud, noisy) a particular song is. 0.0 represents the lowest possible energy song.

Key is a categorical variable that estimates the overall key of a track. The integers map onto standard Pitch Class notation, with 0 representing C, 1 representing C sharp/D flat, etc. If no key is detected, this variable is coded as -1. Loudness refers to the decibel level of the track as averaged across the entire track. They range (approx.) from -60 to 0 dB.
Mode refers to a particular song's modality, whether it is major or minor. Major is coded as 1, while minor is coded as 0.
Speechiness refers to the presence of spoken words on a track. The more speechlike (ex. Like poetry, audio book, talk show) the higher the value for speechiness on a scale of 0.0-1.0.
Acousticness refers to the probability that the track is acoustic, with 0 being definitely not acoustic and 1 being definitely acoustic.
Instrumentalness refers to the probability of vocals on the song, with 0 being there are definitely vocals and 1 being there are definitely no vocals.
Liveness refers to the probability a live audience was present at the time of recording the song.
Valence refers to the relative "positiveness" of a song on a scale from 0 to 1, with 0 being sad-sounding and 1 being happy-sounding.
Tempo refers to the overall average beats per minute (BPM) of the song.
Duration_ms refers to the length of the song in milliseconds.

Section 3. Glimpse of data

```
glimpse(spotify_songs)
```

```
## 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...
```