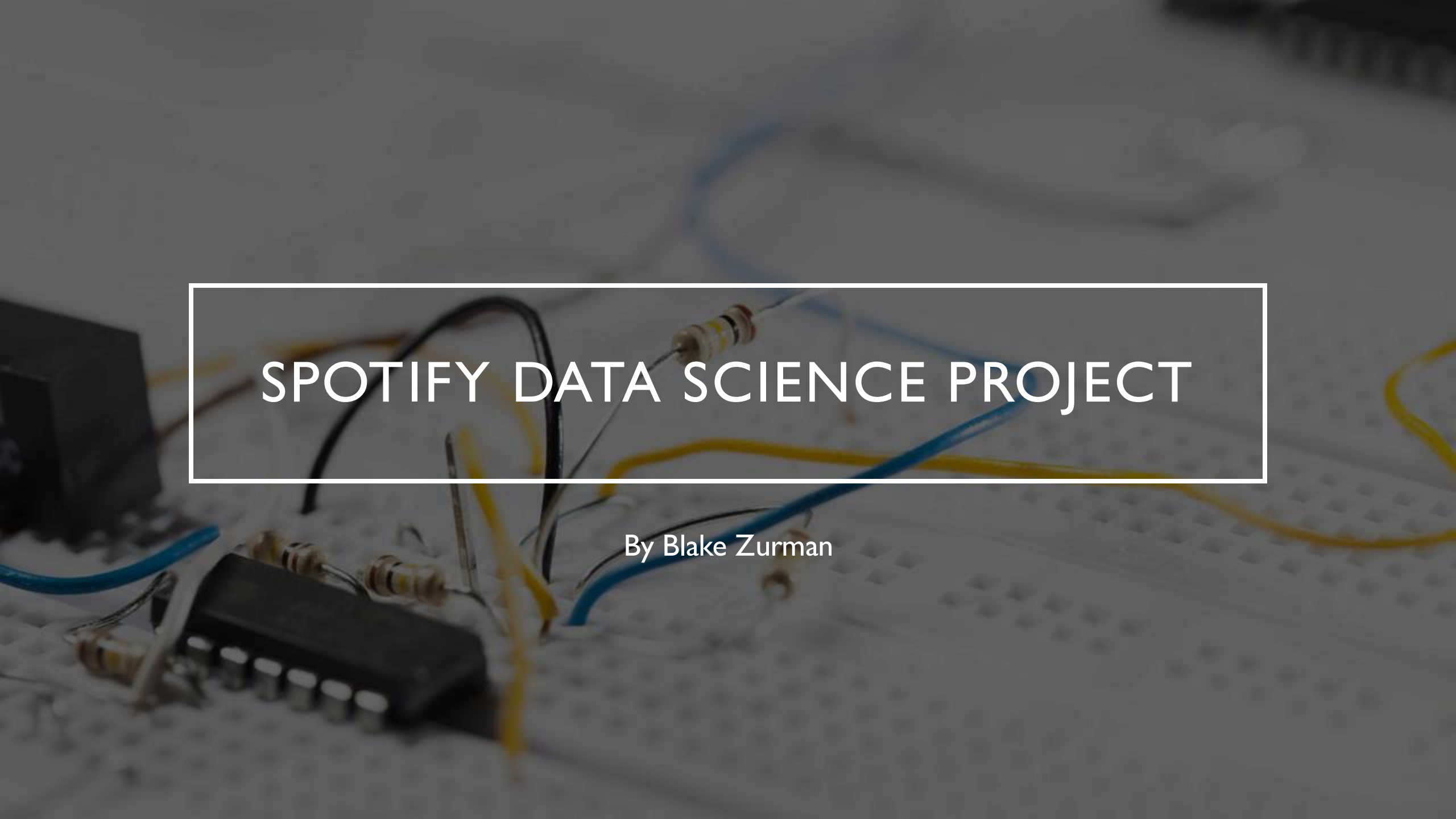


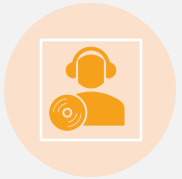
SPOTIFY DATA SCIENCE PROJECT

By Blake Zurman

GOAL OF THE PROJECT

FIND A DATA SET THAT CAN BE USED FOR SONG ANALYSIS, AND A SONG RECOMMENDATION APPLICATION.

DEMONSTRATE DATA CLEANING AND PREPARATION SKILLS ON THE DATA SET.



CONDUCT AN ANALYSIS OF THE DATA TO GAIN INSIGHTS ON MUSICAL DATA.

DISPLAY RESULTS OF THE ANALYSIS TO PROVIDE EASY INTERPRETATION

USED THE CLEANED DATA AND INSIGHTS FROM THE ANALYSIS TO CREATE A WORKING SONG RECOMMENDATION *SHINY* APPLICATION.

THE DATASET

30,000 Spotify songs

The set includes the following attributes: Track name, Artist, Album, Genre, Subgenre, Popularity, Length and Release Date.

The other attributes relating to musicality include: Danceability, Energy, Liveliness, Tempo, Acousticness, Speechiness, Loudness, Instrumentalness, valence, Key, and Mode.

CLEANING

Step 1: Remove the "ID Attribute

- For analysis, this attribute is not necessary.

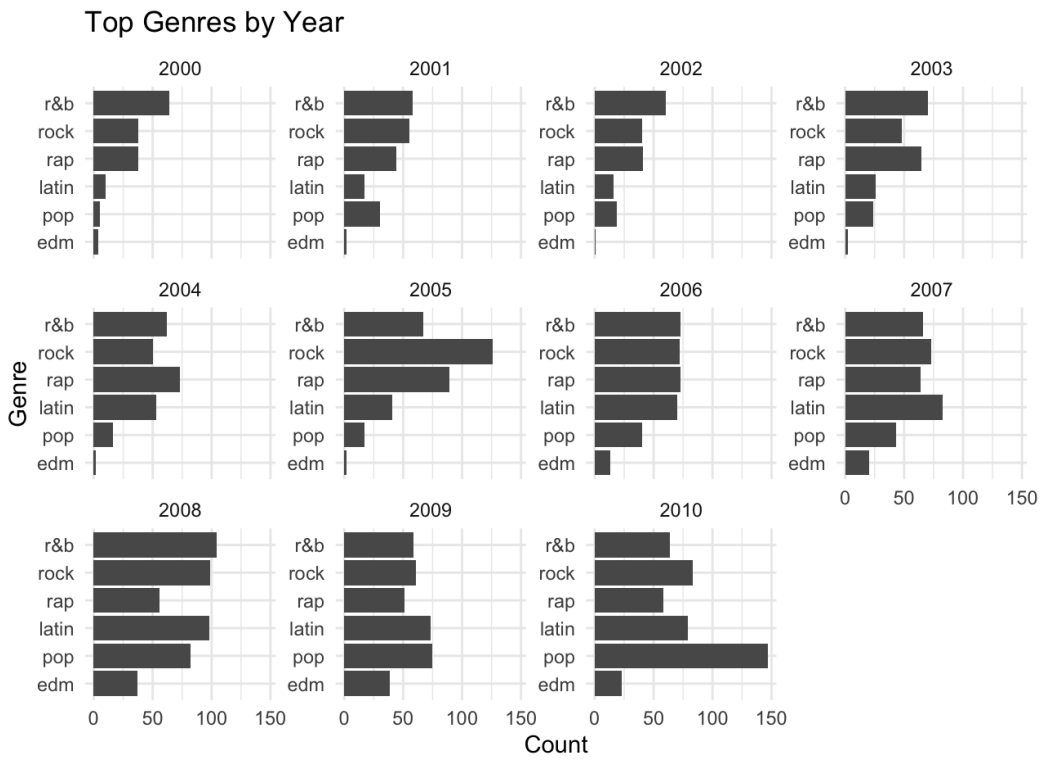
Step 2: Removing Duplicates

- For this data set, duplicate songs occurred because of the way the data set is constructed. Many popular playlists of different genres from Spotify were combined to construct the CSV, but because of human interpretation of what a genre is, genre overlap occurred. This allowed for duplicate songs with different genres.

Step 3: Extract year

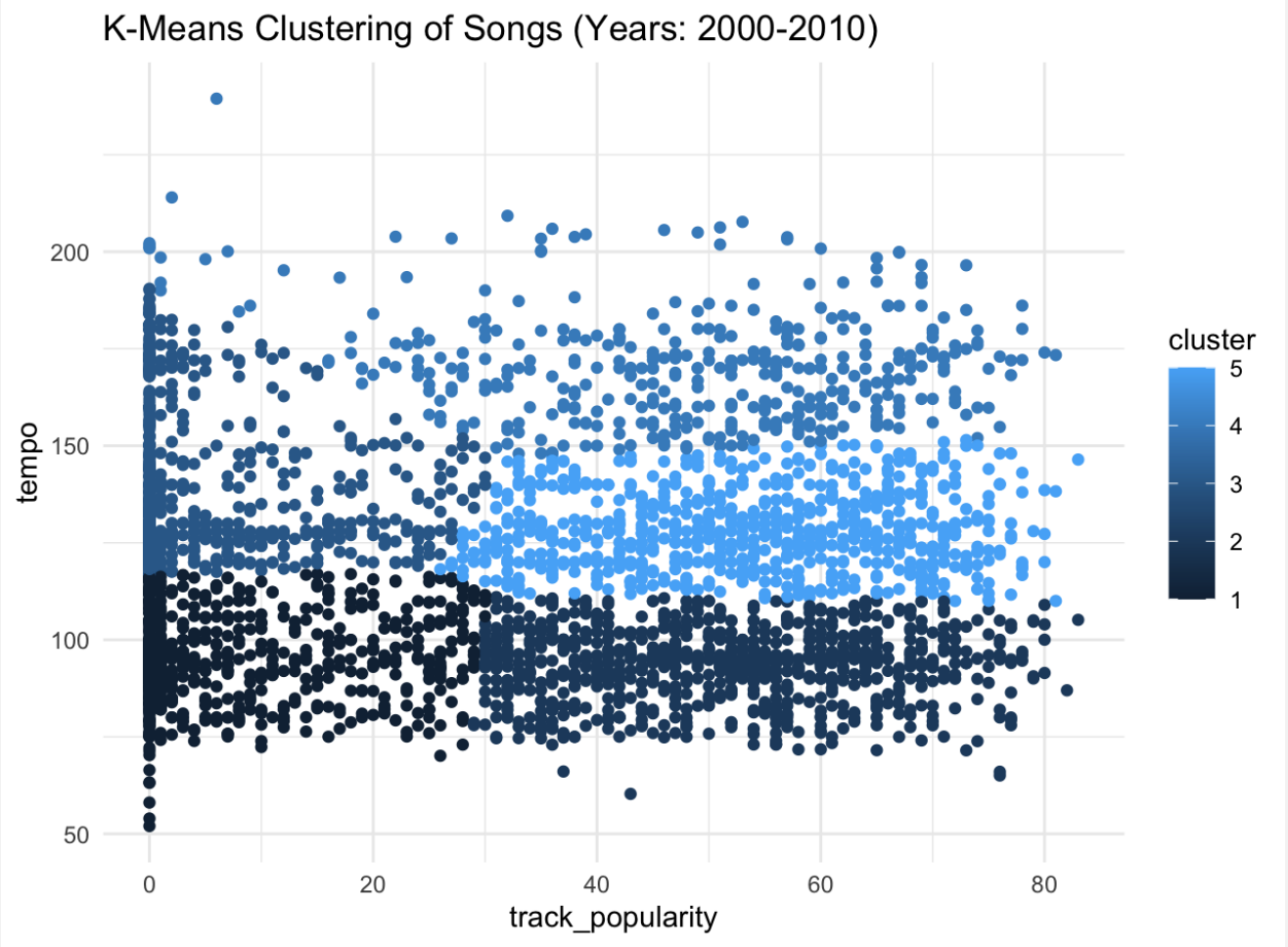
- For analysis, I wanted to focus on specific years and decades, so I made a new column for just release year.

TOP GENRES 2000'S

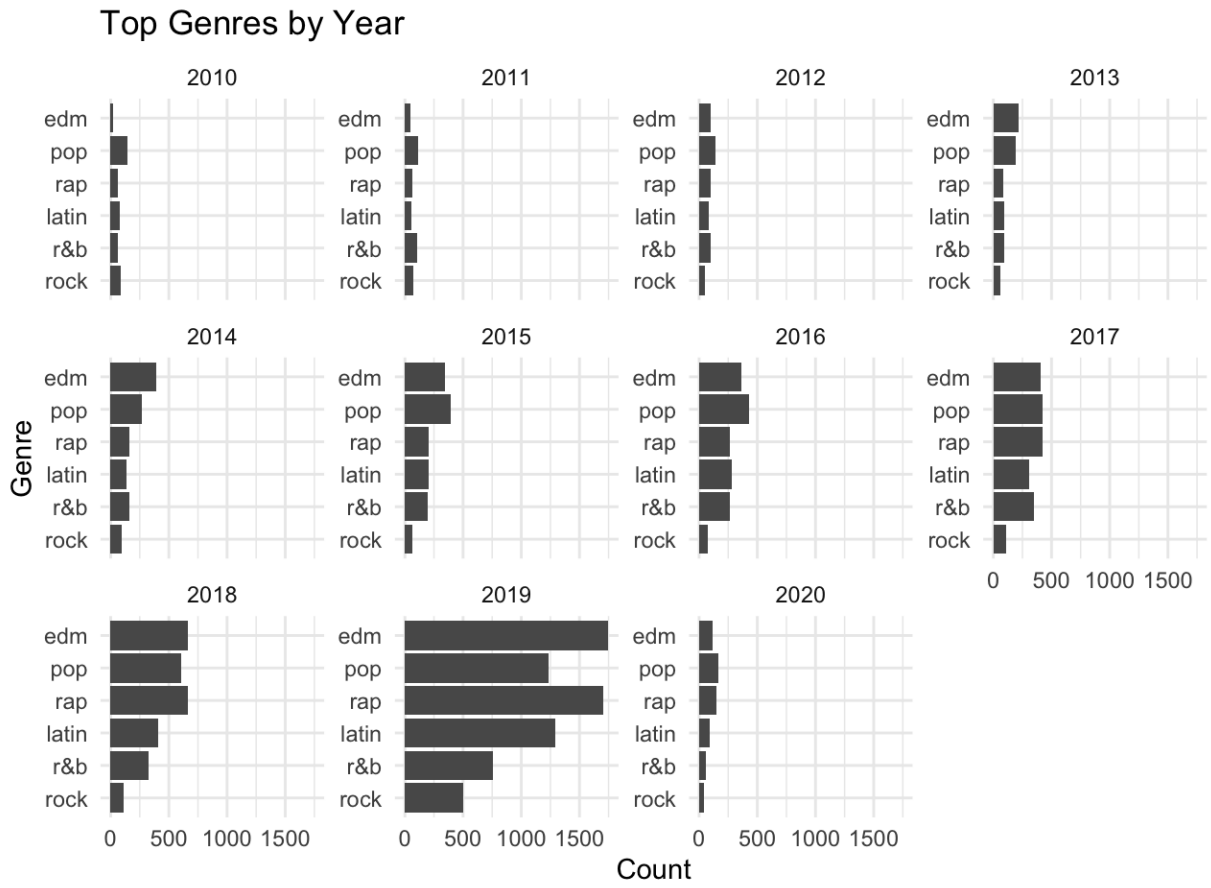


K MEANS CLUSTERING

- 5 clusters of songs gave the best cluster for this data set.
- The model uses nearly all the numeric attributes but is presented in 2D for clarity.
- The colors represent songs that were similar enough to be considered a "group" or a cluster.
- The model received a moderate silhouette score of 0.4104227, meaning the model is ok at clustering the data.

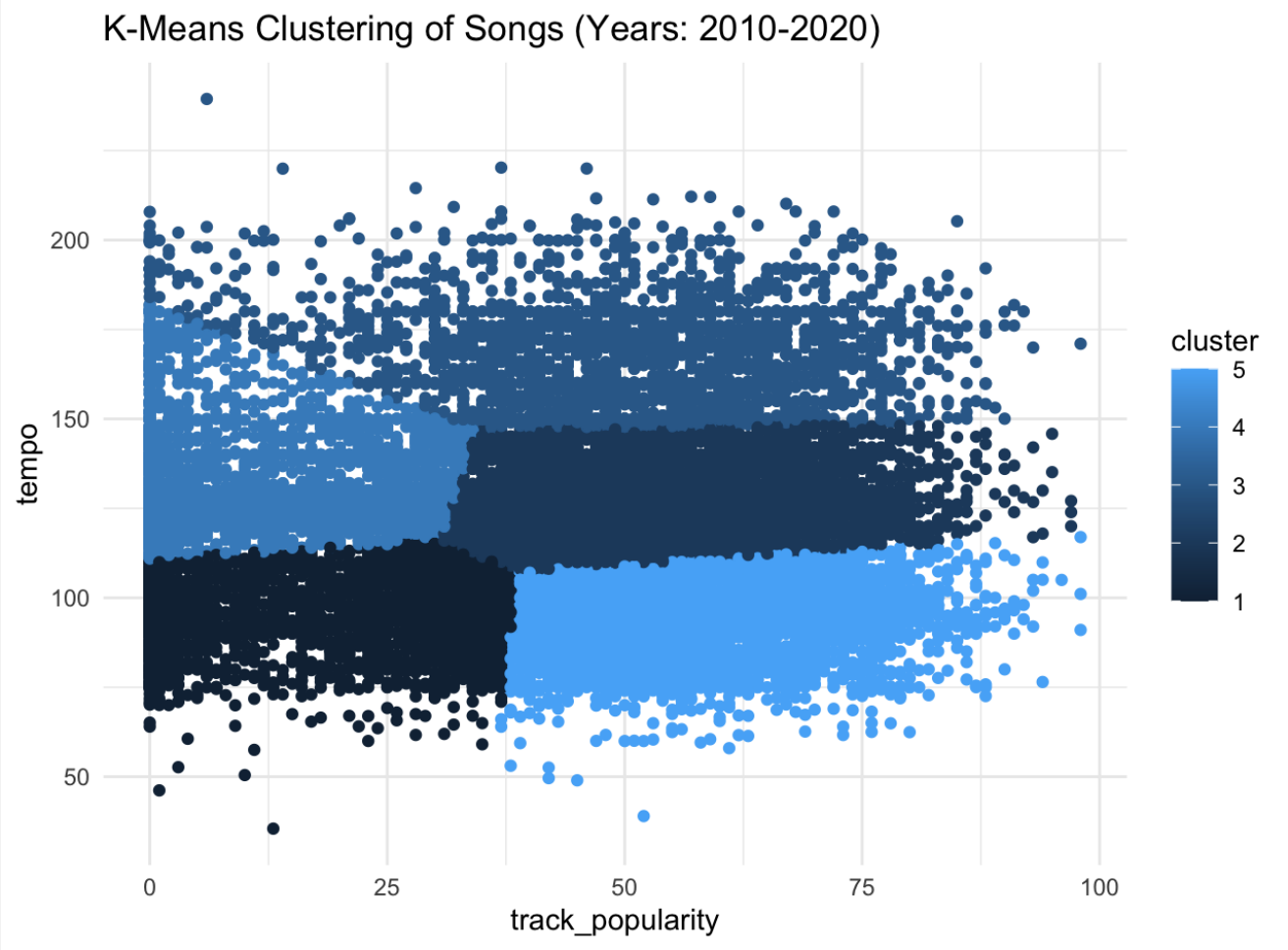


TOP GENRES 2010'S



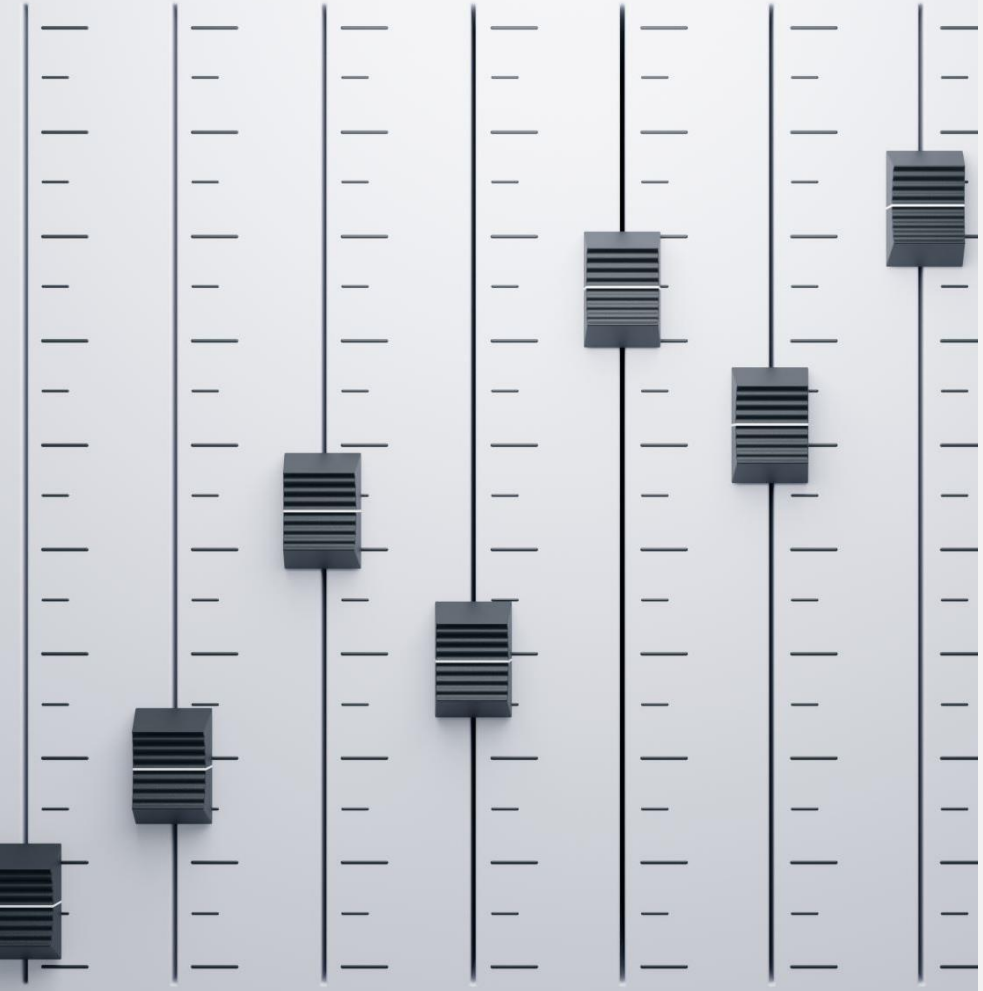
K MEANS CLUSTERING

- 5 clusters of songs gave the best cluster for this data set.
- The model uses nearly all the numeric attributes but is presented in 2D for clarity.
- The colors represent songs that were similar enough to be considered a "group" or a cluster.
- The model received a moderate silhouette score of 0.3686404, meaning the model is still ok at clustering the data.



SVM

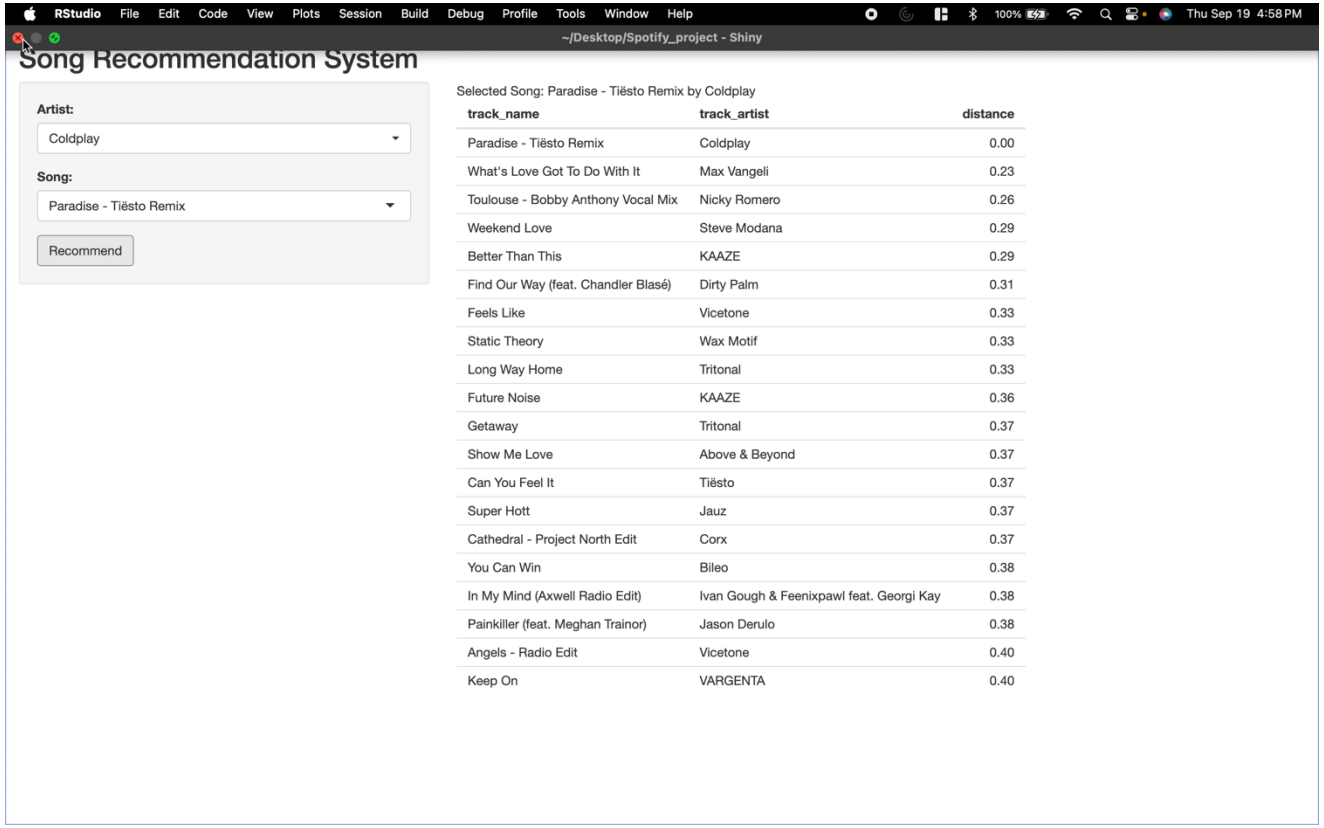
- The goal of the SVM is to predict track genre based on its musical attributes.
- I used SVM because I wanted to do classification, and unsupervised machine learning.
- Overall Statistics:
  - Accuracy: 0.472 (47.2% of predictions were correct)
  - 95% Confidence Interval: (0.4599, 0.4842)
  - No Information Rate (NIR): 0.184 (the accuracy of the most frequent class)
  - P-Value [Acc > NIR]: < 2.2e-16 (indicates that your model is significantly better than random guessing)
  - Kappa: 0.3655 (a measure of agreement between predicted and actual classifications, with values closer to 1 indicating better performance)
  - McNemar's Test P-Value: 4.348e-10 (tests for the difference between two classifiers, indicating significant differences)



SONG RECOMMENDATION

- Based on my analysis of the data set, I decided to use a simple *Euclidean Distance* algorithm to write the logic for the application.
- To simplify, this involves finding the "closest distance" to the selected song to make recommendations.
- I found that the P values of all attributes contributed something to my models, except for the "key" of the song.
- The key of the song describes the scale of musical notes that the song uses.
- This statistical insignificance is to be expected because the key of a song does not typically impact any of the musical attributes in this data set.

DEMO



track_name	track_artist	distance
Paradise - Tiësto Remix	Coldplay	0.00
What's Love Got To Do With It	Max Vangeli	0.23
Toulouse - Bobby Anthony Vocal Mix	Nicky Romero	0.26
Weekend Love	Steve Modana	0.29
Better Than This	KAAZE	0.29
Find Our Way (feat. Chandler Blasé)	Dirty Palm	0.31
Feels Like	Vicetone	0.33
Static Theory	Wax Motif	0.33
Long Way Home	Tritonal	0.33
Future Noise	KAAZE	0.36
Getaway	Tritonal	0.37
Show Me Love	Above & Beyond	0.37
Can You Feel It	Tiësto	0.37
Super Hott	Jauz	0.37
Cathedral - Project North Edit	Corx	0.37
You Can Win	Bileo	0.38
In My Mind (Axwell Radio Edit)	Ivan Gough & Feenixpawl feat. Georgi Kay	0.38
Painkiller (feat. Meghan Trainor)	Jason Derulo	0.38
Angels - Radio Edit	Vicetone	0.40
Keep On	VARGENTA	0.40

CONCLUSION

I am a musician, DJ, and music lover. I believe music is perhaps the most nuanced thing humans can experience, so I knew that a predictive data science project would be a difficult task. I began with lots of optimism around discovering what attracts popularity and listeners. Through data science techniques, I learned about the trends of music through my lifetime, top artists, and the most popular genres. Ultimately, I learned that even when optimizing the hard musical attributes provided in this data set to train an algorithm, music is more than that. It's the song that brings you back to that vacation, the artist whose voice makes your hair stand up, and lyrics that remind you how a loved one talks to you. While machines can understand what it is that we like about music, and even give recommendations, they may never understand the *why*.