STAT199: PROJECT YAAK

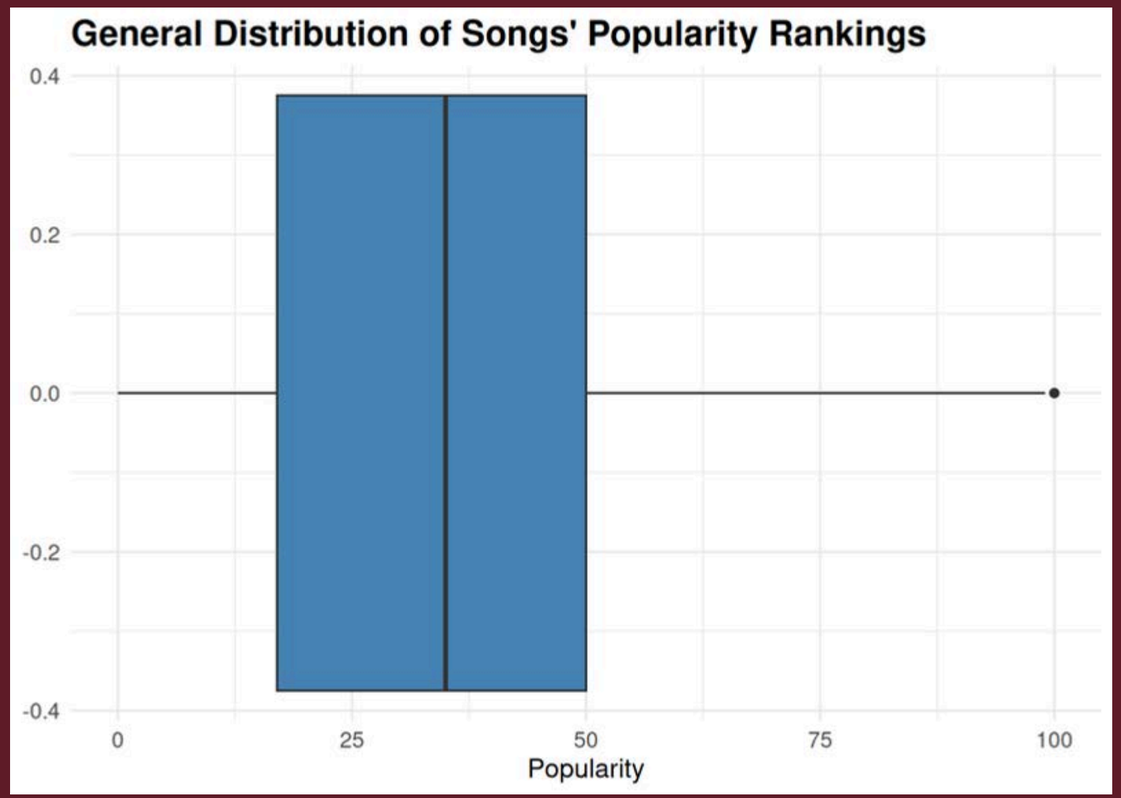
DECODING SPOTIFY: WHAT MAKES MUSIC POPULAR?

OUR DATASET

- KAGGLE SPOTIFY TRACKS DATASET
- 125 GENRES
- INFORMATIONAL DATA - “POPULARITY.”
- MUSICAL DATA - “`DANCEABILITY.”

INITIAL FINDINGS

- RIGHT SKEW
- POPULARITY IS NOT EVENLY DISTRIBUTED



HOW DO DIFFERENT AUDIO FEATURES RELATE TO SONGS' POPULARITY ON SPOTIFY, AND HOW DO THESE FEATURES DIFFER BETWEEN THE MOST AND LEAST POPULAR SONG GENRES?

METHODS

1. Data Cleaning

- Represented mode as a string rather than numerical
- Converted the duration values from ms into min
- Cleaned column names
- Removed unnecessary columns
- Removed any rows with NA values

2. Summary Statistics

- Calculated averages for all song features in the dataset
- Created a new table with these values

3. Creating Groups

- Top 20% versus bottom 5% songs
- Top 5 vs bottom 5 genres

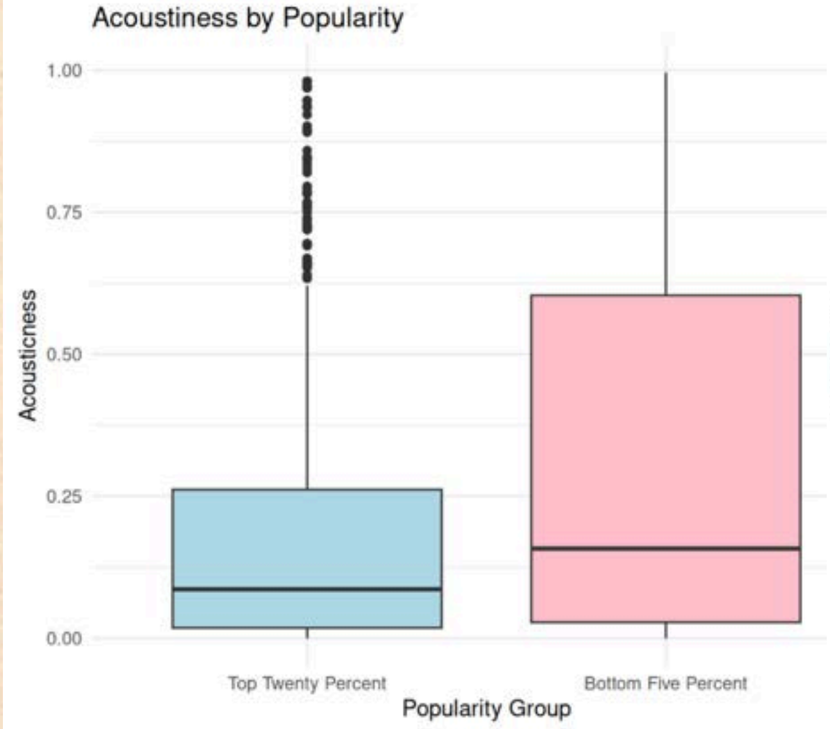
ANALYSIS:

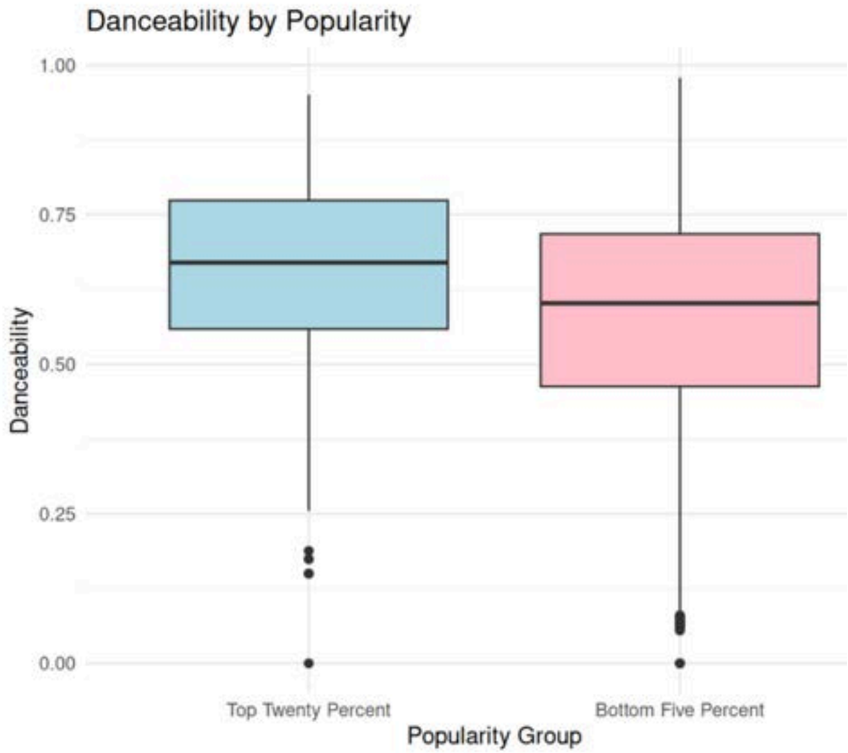
TOP 20% VERSUS BOTTOM 5% OF SONGS

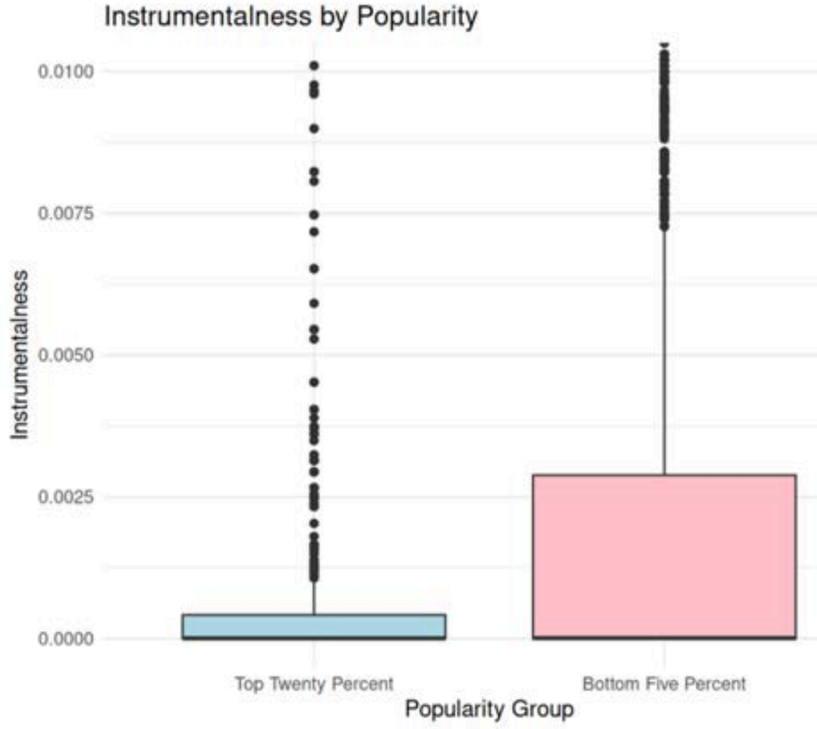
```
$ track_name       <chr> "Hold On", "93 Million Miles", "Unlonely", "Bella Lun…
$ artists          <chr> "Chord Overstreet", "Jason Mraz", "Jason Mraz", "Jaso…
$ album_name       <chr> "Hold On", "Coffee Moment", "Human - Best Adult Pop T…
$ popularity       <dbl> 82, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ duration_ms      <dbl> 198853, 216386, 231266, 302346, 131760, 273653, 13176…
$ explicit         <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS…
$ danceability     <dbl> 0.618, 0.572, 0.796, 0.755, 0.620, 0.633, 0.620, 0.62…
$ energy           <dbl> 0.4430, 0.4540, 0.6670, 0.4540, 0.3090, 0.4290, 0.309…
$ key              <dbl> 2, 3, 5, 9, 5, 4, 5, 5, 5, 5, 6, 0, 0, 0, 6, 2, 10, 6…
$ loudness         <dbl> -9.681, -10.286, -4.831, -9.609, -9.209, -6.784, -9.2…
$ mode             <chr> "Major", "Major", "Minor", "Minor", "Major", "Minor",…
$ speechiness      <dbl> 0.0526, 0.0258, 0.0392, 0.0352, 0.0495, 0.0381, 0.049…
$ acousticness     <dbl> 0.4690, 0.4770, 0.3810, 0.7570, 0.7880, 0.0444, 0.788…
$ instrumentalness <dbl> 0.00e+00, 1.37e-05, 0.00e+00, 0.00e+00, 0.00e+00, 0.0…
$ liveness         <dbl> 0.0829, 0.0974, 0.2210, 0.2360, 0.1460, 0.1320, 0.146…
$ valence          <dbl> 0.167, 0.515, 0.754, 0.330, 0.664, 0.520, 0.664, 0.66…
$ tempo            <dbl> 119.949, 140.182, 97.988, 120.060, 145.363, 143.793, …
$ time_signature   <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,…
$ track_genre      <chr> "acoustic", "acoustic", "acoustic", "acoustic", "acou…
$ duration_min     <dbl> 3.314217, 3.606433, 3.854433, 5.039100, 2.196000, 4.5…
$ pop_group        <fct> Top Twenty Percent, Bottom Five Percent, Bottom Five …
```

ANALYSIS:

TOP 20% VERSUS BOTTOM 5% OF SONGS



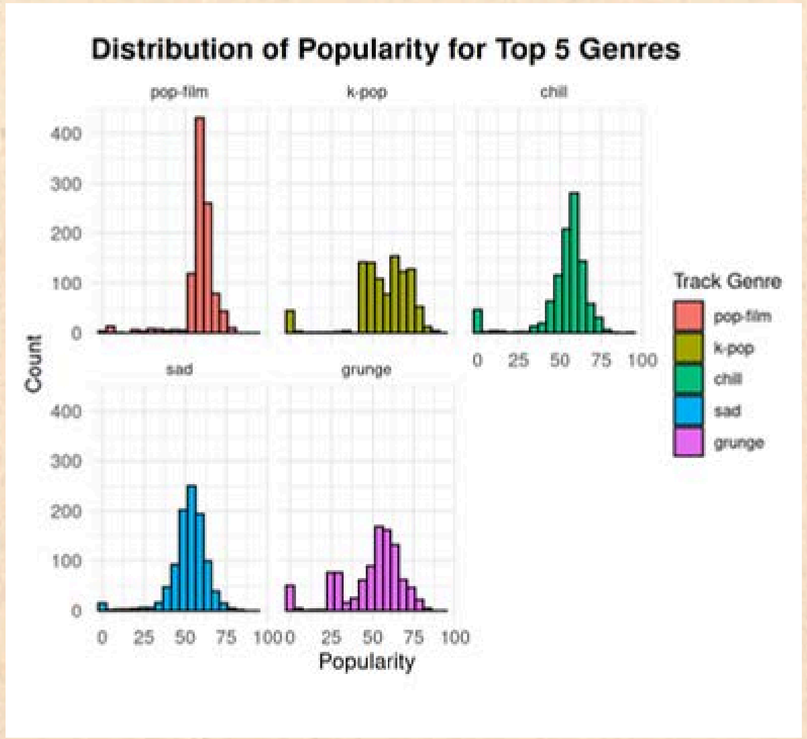


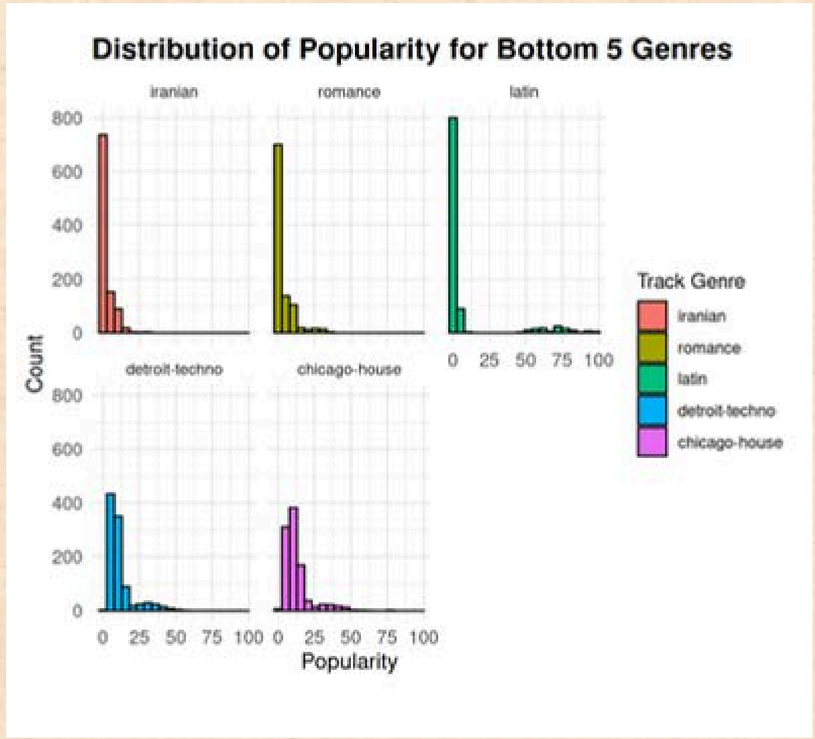


Popular songs tend to be more vocal-driven, slightly more danceable, and less acoustic.

ANALYSIS:

TOP 5 VERSUS BOTTOM 5 GENRES

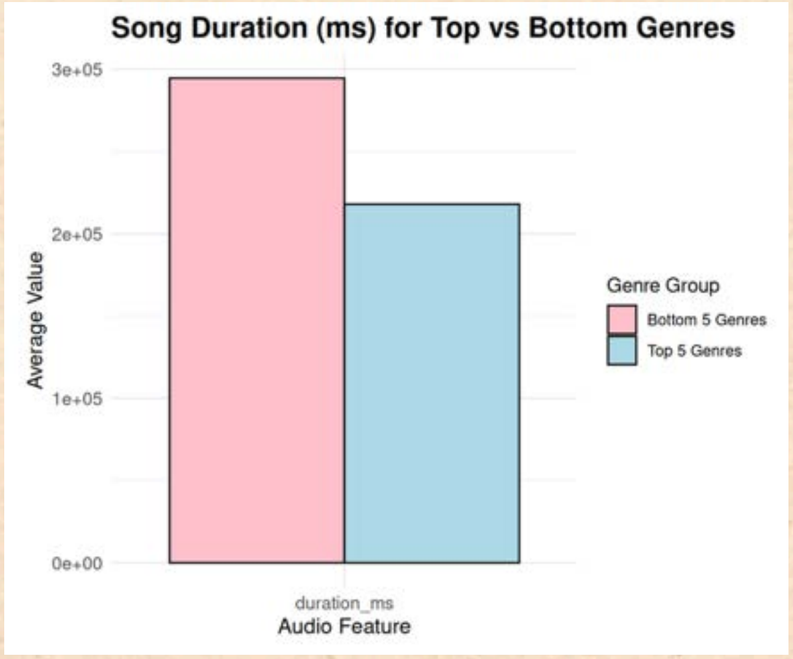


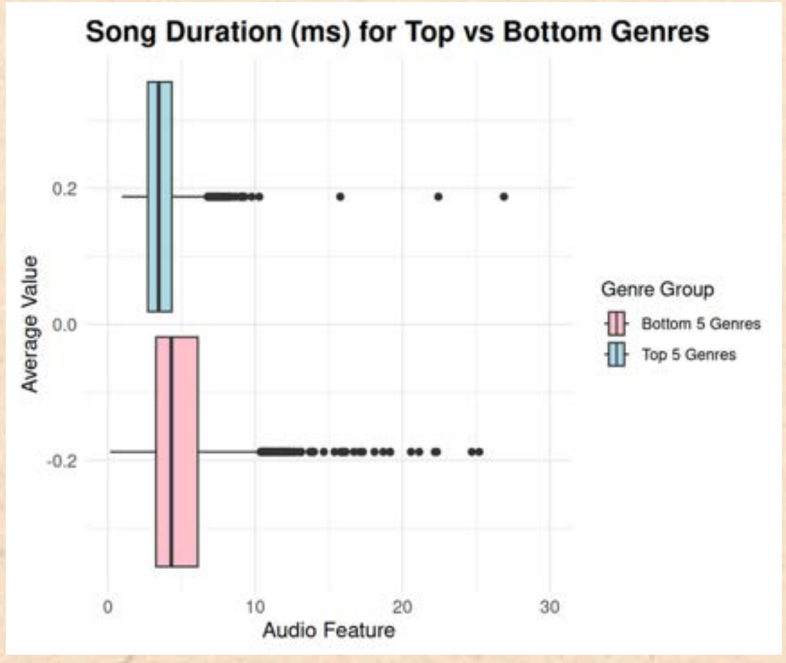


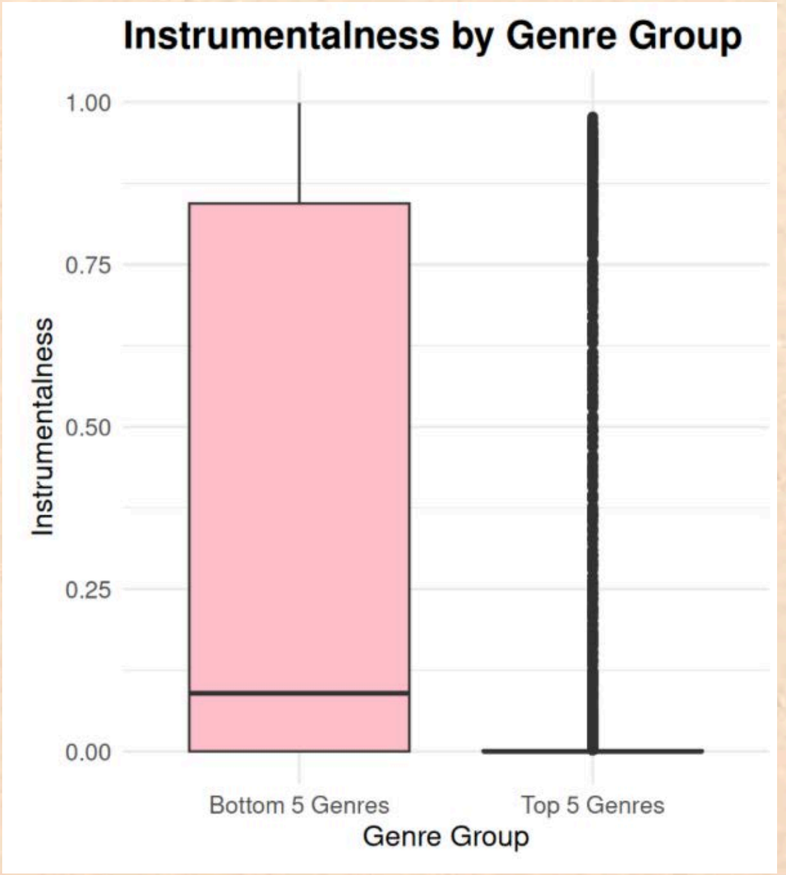

```
$ track_genre     <chr> "pop-film", "k-pop", "chill", "sad", "grunge"
$ avg_popularity <dbl> 59.28300, 56.95295, 53.65100, 52.37900, 49.59400
Rows: 5
Columns: 2
$ track_genre     <chr> "iranian", "romance", "latin", "detroit-techno", "chicago-house"
$ avg_popularity <dbl> 2.210, 3.245, 8.297, 11.174, 12.339
```

ANALYSIS:

TOP 5 VERSUS BOTTOM 5 GENRES







Popular genres tend to be more vocal-driven and shorter in length

DISCUSSION TAKEAWAYS

Traits of Popular Songs on Spotify

- More vocal-driven, less instrumental
- Slightly higher danceability

Traits of Popular Genres on Spotify

- More vocal-driven, less instrumental
- Shorter average song length

LIMITATIONS

- Spotify's popularity is a platform-generated score based on streaming count, not a direct measure of musical quality
- The audio features are Spotify-generated measures, so they may reflect the platform's algorithm bias
- Factors like market, genre tagging, and recency may affect the results
- These findings show associations, NOT proof of causation, since the variables used are platform-generated

THANK YOU