**Summary**

**Data Sources:** Spotify API & Spotify Charts

* **Data Source**: This data set comes from external data sources. I pulled the weekly Top 200 songs played on Spotify for the most recent weeks and eight countries from Spotify Charts. Using these Top 200 listings, I used the Spotify API to pull Spotify’s behind-the-scenes *audio feature* data for each individual song.
* **Data Collection**: Both the Spotify Charts & Spotify API data is all proprietary data owned and generated by Spotify, made available to the public via the internet and free, open-source API data. (**Note**: *The Top 200 data downloaded from Spotify Charts did not include within each CSV file the specific Week, Country, or Country Code for each list. I added these myself in Python to ensure clarity for the analysis.)*
* **Data Contents:**
  + Spotify Charts: This data contains a list of the Top 200 songs per week played in each country. This list includes each songs name, artist, number of times played, weekly rank, and Spotify URL.
  + Spotify API: The data called through Spotify’s API includes track information (song title, album title, etc.) and track audio features (Spotify’s unique classifications for songs which help Spotify’s algorithms generate suggested songs for each user) including Acoustiness, Danceability, Instrumentalness, etc.
* **Resources**:
  + [Spotify Charts](https://spotifycharts.com/regional)
  + [Spotify API](https://developer.spotify.com/documentation/web-api/)
  + [Spotify API Reference Notes](https://developer.spotify.com/documentation/web-api/reference/#endpoint-get-several-audio-features)

**Choosing the Data**

I chose this data because music has always fascinated me and I wanted a chance to learn how to call data through an API. When I learned about Spotify’s unique (and strange) classification system for each song, I wanted a chance to analyze commonalities between audio features. Analyzing data for eight countries over ten weeks should also give me a chance to see whether location, culture, or language play parts in how users choose their favorite songs.

**Data Profile**

**Data Cleaning & Consistency**

* After merging all 80 files from Spotify Charts (8 countries x 10 weeks = 80 csv files) and adding the audio features data for each individual song, I had a file containing 16813 rows of data and 23 columns.
* **Null Values** – 251 rows of data did not have audio features data from the Spotify API. After examining these rows, I realized it was only 26 unique songs over the 10-week period. I removed these 251 rows from the dataset entirely, rather than creating a new API call for 26 songs out of 16,000.
* **New Columns** – In order to make sure I could identify which country in which each song was popular, I added three new columns to each initial file from Spotify Charts – Week, Country, and Country Code.
* **Derived Columns** – In order to streamline the API request for each song’s audio features, I created a new column titled Spotify\_ID, which isolated each song’s unique 22-character identifier from the Spotify URL.
* **Duplicate Values** – The Spotify API generated 840 duplicates because it somehow called down additional rows of audio feature data where specific songs had been released on multiple albums (for instance, a studio album and an compilation album) These duplicates have also been removed..
* **Fixing Mixed-Type Columns** – During the merge of the Top 200 data sets and the audio features data, the Position column was converted to a float64 variable to account for NaN values. After removing the 251 rows without audio features, I reformatted the Position column as int64.

**Introductory Descriptive Statistics**

**Rows**: 15,722

**Columns**: 29

**Total Record Count**: 455938

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Measures** | **Position** | **Streams** | **Length** | **Popularity** | **Acousticness** |
| count | 15722 | 15722 | 15722 | 15722 | 15722 |
| mean | 100.205953 | 872307.594 | 209003.642 | 73.3619132 | 0.279587776 |
| std | 57.6709491 | 1242672.83 | 51078.0445 | 12.4507725 | 0.257717512 |
| min | 1 | 21069 | 52061 | 30 | 2.55E-05 |
| 25% | 50 | 247791.25 | 173746 | 64 | 0.0497 |
| 50% | 100 | 417039 | 203417 | 74 | 0.198 |
| 75% | 150 | 1013753.75 | 240050 | 83 | 0.454 |
| max | 200 | 19747340 | 690731 | 97 | 0.989 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Measures** | **Danceability** | **Energy** | **Instrumentalness** | **Liveness** | **Loudness** |
| count | 15722 | 15722 | 15722 | 15722 | 15722 |
| mean | 0.648508396 | 0.64961633 | 0.00988171 | 0.18087185 | -6.2631556 |
| std | 0.138427241 | 0.17130059 | 0.072315253 | 0.13778567 | 2.40323896 |
| min | 0.161 | 0.0474 | 0 | 0.0202 | -18.064 |
| 25% | 0.555 | 0.525 | 0 | 0.094 | -7.556 |
| 50% | 0.662 | 0.664 | 0 | 0.123 | -6.008 |
| 75% | 0.75 | 0.784 | 1.77E-05 | 0.235 | -4.622 |
| max | 0.971 | 0.989 | 0.927 | 0.977 | 1.906 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Measures** | **Speechiness** | **Tempo** | **Time\_Signature** | **Valence** | **Duration** |
| count | 15722 | 15722 | 15722 | 15722 | 15722 |
| mean | 0.092182947 | 122.707857 | 3.973158631 | 0.53218681 | 208999.492 |
| std | 0.091428767 | 28.6633395 | 0.261698851 | 0.22689392 | 51074.1348 |
| min | 0.0232 | 48.718 | 1 | 0.0348 | 52062 |
| 25% | 0.0364 | 98.951 | 4 | 0.352 | 173747 |
| 50% | 0.0549 | 123.028 | 4 | 0.531 | 203418 |
| 75% | 0.106 | 142.592 | 4 | 0.707 | 240051 |
| max | 0.879 | 208.323 | 5 | 0.978 | 690732 |

**Track Information**

Spotify keeps meticulous records for each song in their library. For this analysis, I’ll be using the following track information:

* **Artists** - The artists who performed the track. Each artist object includes a link in href to more detailed information about the artist.
* **Album** - The album on which the track appears. The album object includes a link in href to full information about the album.
* **Track Name** - The name of the track.
* **Available Markets** - A list of the countries in which the track can be played, identified by their ISO 3166-1 alpha-2 code.
* **Explicit** - Whether or not the track has explicit lyrics (true = yes it does; false = no it does not OR unknown). (Boolean)
* **Track ID** - The unique, 22-character Spotify ID for the track. (String)
* **Popularity** - The popularity of the track. The value will be between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are. Generally speaking, songs that are being played a lot now will have a higher popularity than songs that were played a lot in the past. Duplicate tracks (e.g. the same track from a single and an album) are rated independently. Artist and album popularity is derived mathematically from track popularity. Note that the popularity value may lag actual popularity by a few days: the value is not updated in real time. (Integer)

**Audio Features**

Spotify uses their Audio Features as an in-house classification system upon which they base their predictive algorithms to recommend new music to their listeners. This classification system includes the following variables:

* **Acousticness** - A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic. (Float)
* Danceability - Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable. (Float)
* **Duration ms** - The duration of the track in milliseconds. (Integer)
* **Energy** - A measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy. (Float)
* **Instrumentalness** - Predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal”. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0. (Float)
* **Key** - The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. (Integer)
* **Liveness** - Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live. (Float)
* **Loudness** - The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db. (Float)
* **Mode** - Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0. (Integer)
* **Popularity** - The popularity of the track. The value will be between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are. Note: When applying track relinking via the market parameter, it is expected to find relinked tracks with popularities that do not match min\_\*, max\_\*and target\_\* popularities. These relinked tracks are accurate replacements for unplayable tracks with the expected popularity scores. Original, non-relinked tracks are available via the linked\_from attribute of the relinked track response. (Float)
* **Speechiness** - Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks. (Float)
* **Tempo** - The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration. (Float)
* **Time Signature** - An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). (Integer)
* **Valence** - A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

**Limitations**

While this data is trustworthy because it’s coming directly from Spotify, it’s important to remember that this *only applies to Spotify*. Spotify is a leader in the music streaming market and their data might be reflective of that entire country, but it’s also possible that another streaming service might attract users with different preferences. For instance, Apple Music might have different songs in their Top 200 list for the same countries in the same week.

**Ethical Concerns**

Because Spotify’s API doesn’t allow me to access any user-specific data, the entirety of this analysis concerns national trends in listening preferences. As such, there are not any substantial ethical concerns.

*Note*: It’s always important to understand that this is analysis is representative of a subset of all of Spotify’s users in each country. Therefore, it would be incorrect to stereotype every citizen of a country according to the data found in this analysis. For example, just because Justin Bieber was featured on a song which held a place in the Top 10 songs played in the United States for ten weeks, that doesn’t mean that everyone in the country loves Justin Bieber. I, for one, am not a fan.

**Key Questions**

* What bands are most popular in each country? Can we identify cultural preferences for popular music in each country?
* Which bands transcend those country preferences and have a global appeal?
* What qualities (audio features) make a song popular on a global level? What qualities make a song popular in each country?
* Do Spotify’s audio features help us identify which songs will be more popular than others?
* Can we find an optimal mix of audio features that is predictive of popular music in each country? Globally?