Advanced Analytics & Dashboard Design
6.1: Sourcing Open Data
Prepared by: Cody Boyd

Data Profile

Data Source:
Kaggle: https://www.kaggle.com/datasets/pepepython/spotify-huge-database-daily-charts-over-3-years?select=Database+to+calculate+popularity.csv

Data Collection Method:
This data was sourced directly from Spotify and is internal data that the company shares freely. I sourced the data from Kaggle, where a group of graduate students studying big data prepared and provided the datasets.

Data Context:

- "Database to Calculate Popularity.csv" includes all the daily entries (8mln+) for the songs which made it to the top 200. Among this data, quite intuitively, you will find the same song being in the charts for more than one day. We then created a popularity score, unique for a given song in each country, which considered the position in the charts and the days it stayed there.
- "Final Database.csv" includes a lot of data for each song. It aggregates the popularity for songs into a single score for each. For each song several variables were retrieved by using Spotify's API (such as artist, country, genre, …)

Data Contents:
This huge dataset contains all the songs in Spotify's Daily Top 200 charts in 35+1 (global) countries around the world for a period of over 3 years (2017-2020).

Data Limitations and Ethics:
The company is a third party that shares and streams music for subscribers from around the world, however, there could be biases from the company as if it gains revenue from ranking music (The creators of the data set made a popularity scale due to the bias nature of Spotify's ranking system).

Data Relevance:
This data appears to meet all the qualifications for this achievement given in the Achievement 6 project brief. I chose this data because I'm passionate about music, and Spotify in particular. I have worked in the entertainment industry for 7 radio stations, working with many Artists, Artist managers, concerts, and concert venues. The aspects of my current and former jobs that I enjoyed the most were my analytic tasks and preparing visualizations, and I've found that I regularly use my Spotify app in my free time to relax and catch vibes. This data seems like a good fit for my interests.

Data Profile

Database to calculate popularity.csv

- Original data contains 9807001 rows and 8 columns.

Final database.csv

- Original data contains 170633 rows and 151 columns.

- Variables Used:
  - Title: Name of a song
  - URI: Unique identifier of a song created by Spotify
  - Country: Global and 34 countries where Spotify operates (Argentina, Australia, Austria, Belgium, Brazil, Canada, Chile, Colombia, Costa Rica, Denmark, Ecuador, Finland, France, Germany, Great Britain, Indonesia, Ireland, Italy, Mexico, Malaysia, Netherlands, New Zealand, Norway, Peru, Philippines, Poland, Portugal, Singapore, Spain, Sweden, Switzerland, Taiwan, Turkey, USA)
  - Popularity: The popularity score calculated considering both the number of days a song stayed in the Top200 and the position it stayed in every day, weighting more the top positions
  - Artist: Name of the songs' artist
  - Album/Single: Whether the song was published as a single or as part of an album or compilation
  - Genre: The predominant genre of an artist according to Spotify's classification
  - Artistfollowers: The number of followers the artist has on Spotify on the 5th of November 2020
  - Explicit: Whether the song is rated as 'Parental Advisory Explicit Content' or not
  - Album: Name of the album the song belongs to
  - Releasedate: Date on which the song was published
  - Track_number: The position of the song on its respective album
  - Track _album: Total songs present in the album
  - Danceability: 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
  - Energy: It is 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. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy
  - Key: The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation. E.g., 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1
  - 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
  - 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
  - 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
  - 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
  - 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
  - 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

  - 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)
  - 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
  - Durationms: The duration of the track in milliseconds
  - Timesignature: 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)
  - Genrenew: The predominant genre of an artist according to our reclassification of Spotify's 1200 genres
  - Dayssincerelease: Number of days passe from the release of the track
  - Releasedafter2017: Dummy equal to 1 if the track was released after 2017
  - Syuzhetnorm: Tone of the lyrics based on the library Syuzhet. The initial result is normalized on the scale: -1 (negative), 0 (neutral), 1 (positive).
  - Angernorm2: Number of words related to anger divided by the total number of words found by the dictionary
  - Anticipationnorm2: Number of words related to anticipation divided by the total number of words found by the dictionary
  - Disgustnorm2: Number of words related to fear divided by the total number of words found by the dictionary
  - Fearnorm2: Number of words related to fear divided by the total number of words found by the dictionary
  - Joynorm2: Number of words related to joy divided by the total number of words found by the dictionary
  - Sadnessnorm2: Number of words related to sadness divided by the total number of words found by the dictionary
  - Surprisenorm2: Number of words related to surprise divided by the total number of words found by the dictionary
  - Trustnorm2: Number of words related to trust divided by the total number of words found by the dictionary
  - Bayes: Tone of the lyrics according to Bayes on the scale: -1 (negative), 0 (neutral), 1 (positive)
  - LDATopic: Topic of the lyrics according to the categories: Love, Thug, Nostalgia, Explore, Fun, Desire, Hope and Celebrate
  - Popumax: The top position reached by a track in the 1401 days we have data for
  - Top50_dummy: A dummy equal to 1 if the top position reached by a song is 50 or higher

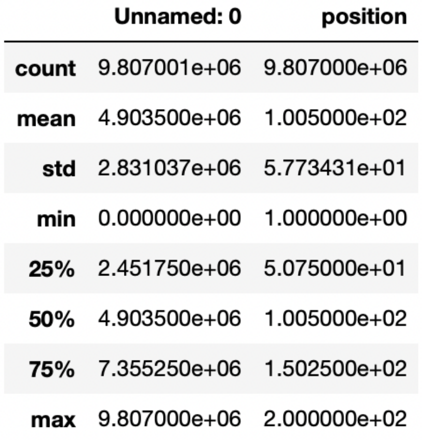
- Consistency Checks:
  - Data types corrected as needed.
    - "Artist_followers" changed to string
    - "Explicit" changed to string
    - "LDA_Topic" changed to string

- Missing Values
  - Column "Artist" contained 15,642 NaN values
    - Created a subset for missing values
  - URI # 9807000 contained missing values for 6 variables, however, I left it alone.

- Duplicate Values
  - No duplicates were present in either data frame.

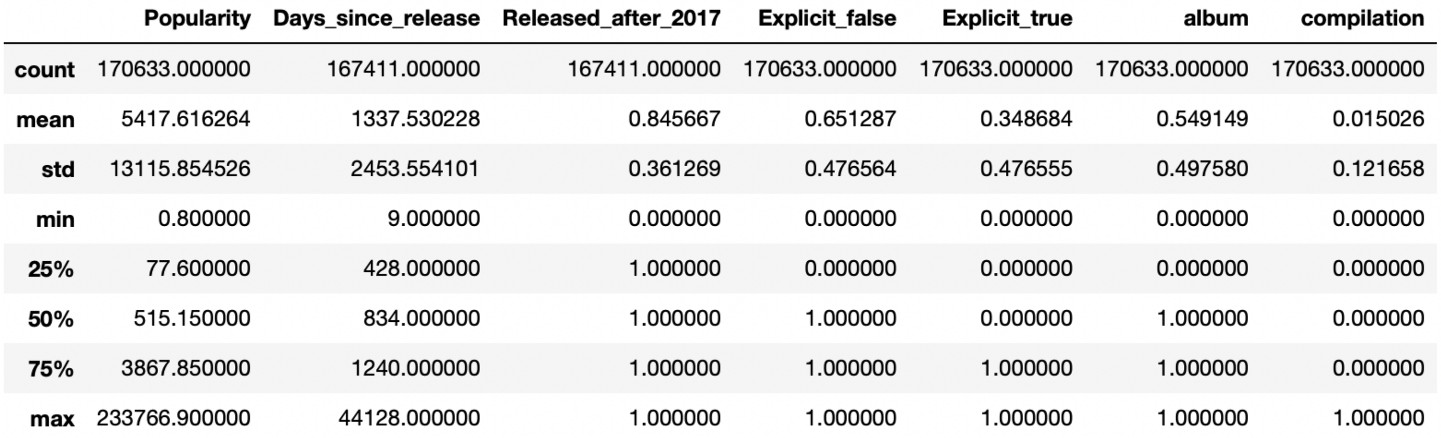
- Descriptive Statistics:

Database to calculate popularity.csv



	Unnamed: 0	position
count	9.807001e+06	9.807000e+06
mean	4.903500e+06	1.005000e+02
std	2.831037e+06	5.773431e+01
min	0.000000e+00	1.000000e+00
25%	2.451750e+06	5.075000e+01
50%	4.903500e+06	1.005000e+02
75%	7.355250e+06	1.502500e+02
max	9.807000e+06	2.000000e+02

Final database.csv



	Popularity	Days_since_release	Released_after_2017	Explicit_false	Explicit_true	album	compilation
count	170633.000000	167411.000000	167411.000000	170633.000000	170633.000000	170633.000000	170633.000000
mean	5417.616264	1337.530228	0.845667	0.651287	0.348684	0.549149	0.015026
std	13115.854526	2453.554101	0.361269	0.476564	0.476555	0.497580	0.121658
min	0.800000	9.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	77.600000	428.000000	1.000000	0.000000	0.000000	0.000000	0.000000
50%	515.150000	834.000000	1.000000	1.000000	0.000000	1.000000	0.000000
75%	3867.850000	1240.000000	1.000000	1.000000	1.000000	1.000000	0.000000
max	233766.900000	44128.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Potential Questions to Explore

1. Do countries share the same top-ranking artists or songs?
2. Is there a correlation between popularity of an artist and being in the top 50 charts?
3. Are there different genres that are most popular per country?
4. Who is the most popular artist in the U.S. currently and by each recorded decade?