SPOTIFY DATA ANALYSIS

Charles Huang and Zuoqi Zhang

GOALS

- What top songs have in common.
- What elements lead to a song ranking highly or becoming popular.
- What features allow a song to stay popular for an extended period of time.
- Given a new song, predict how popular it will be. (ranking, streams)

DATA SELECTION

- Kaggle
  - January 2017 to January 2018
  - 200 most listened to songs on Spotify for each day per region
  - 53 countries
  - 6,629 artists
  - 18,598 songs
  - Position, track name, artist, streams, URL, date, region
- Spotify Web API and spotipy Python library
  - 13 Audio features:
    - Acousticness, danceability, duration_ms, energy, instrumentalness, key, liveness, loudness, mode, speechiness, tempo, time_signature, valence
- Multiple versions of songs kept as is
  - The original song and a remix are considered to be two separate songs
- Only consider main artist of each song
  - Does not include featured artists
- We chose to use the US data, but our model should be able to use any region's data
  - 1,967 songs

AUDIO FEATURE DEFINITIONS

Audio Features Object

KEY	VALUE TYPE	VALUE DESCRIPTION
acousticness	float	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
analysis_url	string	An HTTP URL to access the full audio analysis of this track. An access token is required to access this data.
danceability	float	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.
duration_ms	int	The duration of the track in milliseconds.
energy	float	Energy 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. 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.
id	string	The Spotify ID for the track.
instrumentalness	float	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.
key	int	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.
liveness	float	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.
loudness	float	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	int	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	float	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.
tempo	float	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.
time_signature	int	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).
track_href	string	A link to the Web API endpoint providing full details of the track.
type	string	The object type: “audio_features”
uri	string	The Spotify URI for the track.
valence	float	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).

ASSUMPTIONS

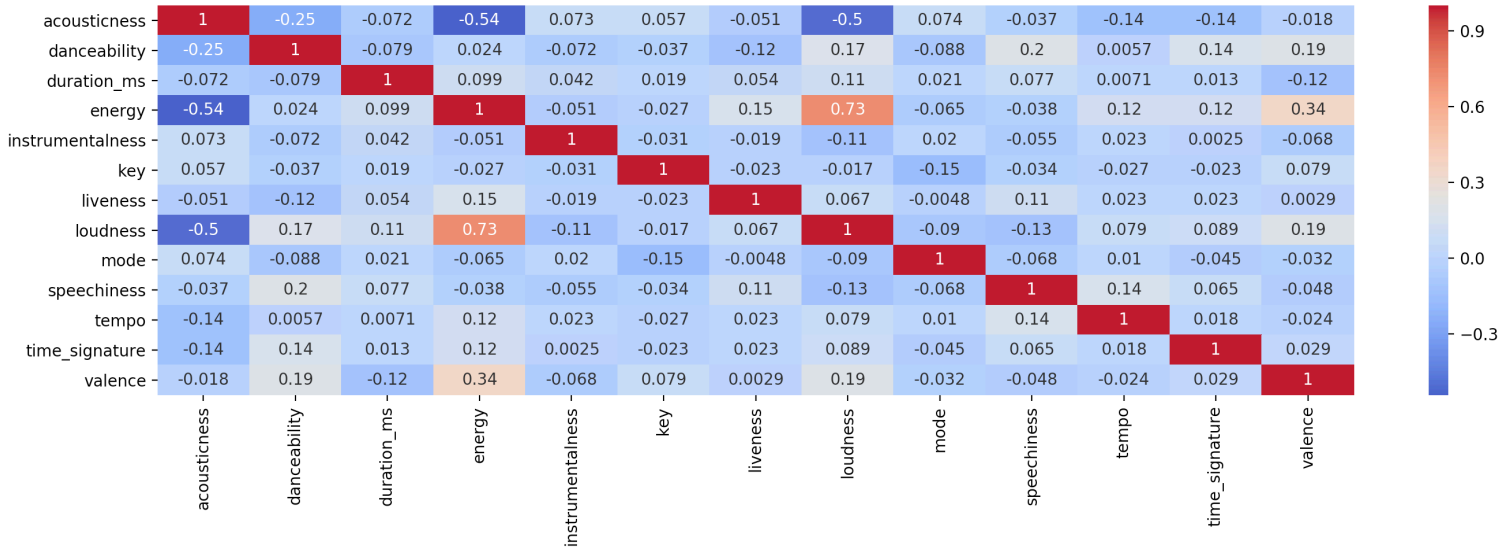
- Artist name and track name greatly impact popularity
  - The more well known they are, the more likely people are to listen to them
- Aim to find features other than these two that may also impact popularity

Rank	Artist	Streams
1	Drake	1250864578
2	Kendrick Lamar	1163890899
3	Post Malone	988811897
4	Lil Uzi Vert	773675118
5	Ed Sheeran	723507889
6	Migos	694024669
7	Future	568147747
8	The Chainsmokers	556698859
9	21 Savage	481196247
10	Khalid	470519202

Rank	Track Name	Streams
1	HUMBLE.	339677217
2	XO TOUR Llif3	316206696
3	Congratulations	285451131
4	Shape of You	282319891
5	Unforgettable	264449000
6	Mask Off	241828211
7	rockstar	238013267
8	Despacito - Remix	235012075
9	iSpy (feat. Lil Yachty)	227087919
10	Location	226224851

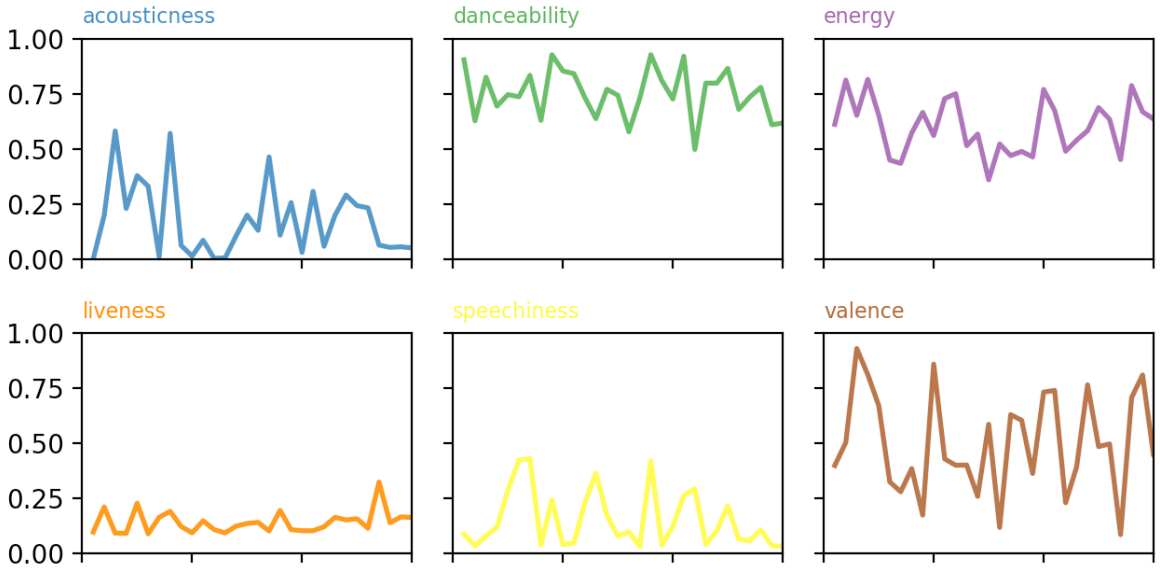
DATA ANALYSIS

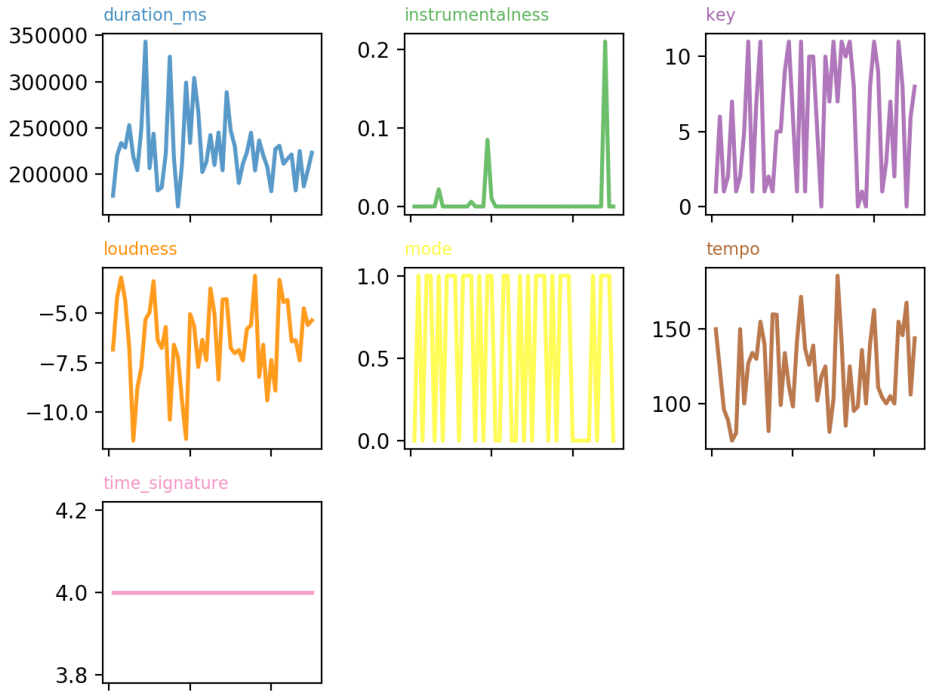
- Correlation between different features
- Analyze features to see which features may be closely related
- Popular songs assumed to have high correlation between popular features
  - For instance, if having high energy is a popular feature, having high loudness most likely also is, but having high acousticness would not be



DATA ANALYSIS CONTINUED

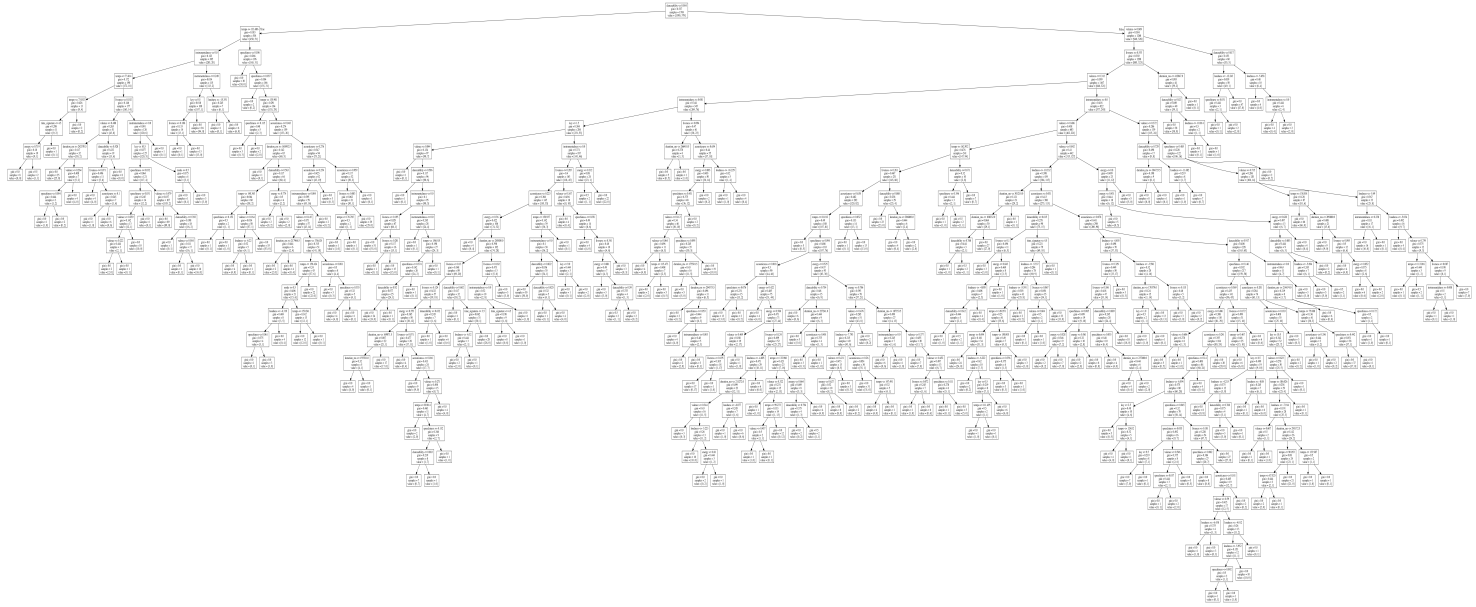
- Trends for each feature among the songs (top 50)
- No clearly visible trends
  - May be due to previously top ranked songs moving down in the rankings





CLASSIFICATION

- Used audio features and total streams per song to see if features can determine popularity
  - Accousticness, danceability, duration_ms, …
- Labels:
  - 0: streams < average -> less popular
  - 1: streams >= average -> more popular
- Used sklearn's implementation of KNN and decision tree
- Split the data:
  - 90% training data
  - 10% testing data
- K Nearest Neigbors:
  - Result: 78% accuracy
- Decision Tree:
  - Result: 72% accuracy
- Chose to use KNN



FUTURE WORK

- Obtain the dates on which the songs were released
  - Release dates may impact popularity since newer songs are typically assumed to be more popular
- Compare results of models with different subsets of features used
  - Find best predictors
- Currently using a binary classification method, but want to switch to a multiclass classification method
  - Top 50, 51~100, 101~150, 151~200...

REFERENCES

- Spotify Web API: https://beta.developer.spotify.com/documentation/web-api/
- Kaggle: https://www.kaggle.com/edumucelli/spotifys-worldwide-daily-song-ranking/data