Data Analytics and Statistical Learning: Final Project

a. Members' names

Yanniode Peri-Okonny
Petra Kumi
Philippe Lessard
Camelia D. Brumar

b. Description of the problem

We want to predict how well ranked would a new song be ranked in a given country based on its several features of the music, including artist, time signature, key, and danceability.

c. Description of the data set (dimensions, names of variables with their description)

There are 2 datasets that we plan to combine. One is Spotify's Worldwide Daily Song Ranking, which has 3.44 million rows and 7 columns, which are listed below. We plan to only consider the point at which songs peak on the charts, which should significantly reduce the number of rows. We also will limit the number of countries we will consider, which will further reduce the number of rows.

Position	Position on the charts
Track Name	The name of the song
Artist	Name of the artist
Streams	Number of streams
URL	URL for the song
Date	The date the data is from
Region	The country

The second dataset is the Spotify Tracks DB, which contains various features for tracks on Spotify, such as energy and danceability. The dataset contains 228,159 tracks. The dataset's description also contains the code used to get the data, which we can use to get more data if we are missing songs. The full list of features is in the table below.

genre	The genre of the song
artist_name	The name of the artist
track_name	The name of the song
track_id	The Spotify id for the song (can be used to find the song on Spotify)
popularity	The position on the chart
acousticness	Confidence of whether the track is acoustic. Float between 0.0 and 1.0
danceability	Estimate of how suitable the track is for dancing. Float between 0.0 and 1.0
duration_ms	The duration of the track in milliseconds
energy	Measures intensity and activity. Float between 0.0 and 1.0
instrumentalness	Guesses if the track has vocals. Float between 0.0 and 1.0. 1.0 means a track does not contain vocals
key	An estimate of the overall key of the track. Maps pitch classes (C, C#, D, etc) to integers
liveness	Guesses if the track was recorded live. Float between 0.0 and 1.0. 1.0 indicates high probability the track is live
loudness	Average loudness of the track in dB. Float with typical range of -60 to 0
mode	Whether the track is major or minor key. Major = 1, minor = 0
speechiness	Represents how exclusively speech-like the recording is. Float between 0.0 and 1.0. Closer to 1.0 represents more speech-like
tempo	An estimate of the overall tempo of the song in BPM.
time_signature	An estimate of the overall time signature of the track
valence	Represents the positiveness of the track. Float between 0.0 and 1.0. Higher values

	represent more positive sounding songs

d. Regression or classification?

Classification; determining whether the song will rise into the Top 5, Top 10, Top 50, Top 100, or Top 200.

e. The methods you plan to try.

The ones we have in mind are K-NN, LDA, QDA and Logistic Regression. We have to figure out if the data has the same variance to decide between LDA and QDA, and if it is Gaussian with respect to the predictors we will use. Also, we will see if K-NN works decently using the quantity of data points we have. If we will have enough time, we will try also using logistic regression and do cross validation to see which of the methods gives us the best accuracy.

f. The error metrics you plan to use and the algorithms for assessing them.

We will use RSS for our error metrics, and we will use cross-validation to decide between using AIC or BIC to assess our error.

g. Comments and/or concerns?

If we have any time left before the final submission, we will create a user interface with scrollable bars for some of the most relevant attributes, so the user can change the value of the corresponding predictors and see how a song with those attributes would be ranked.

Dataset link:
https://www.kaggle.com/zaheenhamidani/ultimate-spotify-tracks-db
https://www.kaggle.com/edumucelli/spotifys-worldwide-daily-song-ranking