Description of dataset: This dataset consists of data on 52,000 songs that were randomly picked from a variety of genres sorted in alphabetic order (a as in "acoustic" to h as in "hiphop"). For the purposes of this analysis, you can assume that the data for one song are independent for data from other songs.

This data is stored in the file "spotify52kData.csv", as follows:
Row 1: Column headers
Row 2-52001: Specific individual songs
Column 1: **songNumber** – the track ID of the song, from 0 to 51999.
Column 2: **artist(s)** – the artist(s) who are credited with creating the song.
Column 3: **album_name** – the name of the album
Column 4: **track_name** – the title of the specific track corresponding to the track ID
Column 5: **popularity** – this is an important metric provided by spotify, an integer from 0 to 100, where a higher number corresponds to a higher number of plays on spotify.
Column 6: **duration** – this is the duration of the song in ms. A ms is a millisecond. There are a thousand milliseconds in a second and 60 seconds in a minute.
Column 7: **explicit** – this is a binary (Boolean) categorical variable. If it is true, the lyrics of the track contain explicit language, e.g. foul language, swear words or otherwise content that some consider to be indecent.
Column 8: **danceability** – this is an audio feature provided by the Spotify API. It tries to quantify how easy it is to dance to the song (presumably capturing tempo and beat), and varies from 0 to 1.
Column 9: **energy** - this is an audio feature provided by the Spotify API. It tries to quantify how "hard" a song goes. Intense songs have more energy, softer/melodic songs lower energy, it varies from 0 to 1
Column 10: **key** – what is the key of the song, from A to G# (mapped to categories 0 to 11).
Column 11: **loudness** – average loudness of a track in dB (decibels)
Column 12: **mode** – this is a binary categorical variable. 1 = song is in major, 0 – song is in minor
Column 13: **speechiness** – quantifies how much of the song is spoken, varying from 0 (fully instrumental songs) to 1 (songs that consist entirely of spoken words).
Column 14: **acousticness** – varies from 0 (song contains exclusively synthesized sounds) to 1 (song features exclusively acoustic instruments like acoustic guitars, pianos or orchestral instruments)
Column 15: **instrumentalness** – basically the inverse of speechiness, varying from 1 (for songs without any vocals) to 0.
Column 16: **liveness** - this is an audio feature provided by the Spotify API. It tries to quantify how likely the recording was live in front of an audience (values close to 1) vs. how likely it was recorded in a studio without a live audience (values close to 0).
Column 17: **valence** - this is an audio feature provided by the Spotify API. It tries to quantify how uplifting a song is. Songs with a positive mood =close to 1 and songs with a negative mood =close to 0
Column 18: **tempo** – speed of the song in beats per minute (BPM)
Column 19: **time_signature** – how many beats there are in a measure (usually 4 or 3)
Column 20: **track_genre** – genre assigned by spotify, e.g. "blues" or "classical"

Note that we did most of the data munging and coding for you already but you *might* still need to handle missing data in some way (e.g. by row-wise removal, element-wise removal, imputation). Also, if there are skewed distributions, extreme values might also have to be handled.