466 Project Proposal

Team: Braeden Alonge, Lucas Summers, Xiuyuan Qiu, Nathan Lim

Dataset Info:

- https://www.kaggle.com/datasets/joebeachcapital/30000-spotify-songs/data
- 10 Spotify-generated features + other metadata for 30k Spotify songs on Kaggle pulled from the Spotify API
- This dataset includes various numeric musical attributes such as track popularity, danceability, energy, loudness, speechiness, tempo, key, loudness, and more.
- It also contains metadata like song name, artist, album, release date, duration, and playlist information (genre, subgenre, etc.)

Questions to Study:

- How do different song features correlate with each other?
- What are the key differences between popular and unpopular songs?
- How have song features evolved over time?
- Can we predict the genre of a song based on its features?
- Do certain artists consistently produce songs with specific features?
- Can we identify common features within specific Spotify playlists?

potential datasets:

- https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success
  - Dataset with a lot of features to predict academic success.
  - lots of features and lots of observations and data preprocessing (outliers + missing data) have already been handled
- https://archive.ics.uci.edu/dataset/240/human+activity+recognition+using+smartphones
  - dataset attempts to classify human activity based on smartphone spatial data
  - Lots of observations and we can use both clustering and classification in our investigation
- https://www.kaggle.com/competitions/instacart-market-basket-analysis
  - competition dataset which makes it intuitive for us to analyze
  - main question is what types of items customers will rebuy and what are the giveaways for that
  - literal metric ton of observations, and is a real dataset!
  - not just a singular dataset (multiple csv files), which might make it more complex but also can make it better for complex analysis if we want.
- https://www.kaggle.com/datasets/insiyeah/musicfeatures
  - Features of 1000 music clips (30s each), categorized by genre (100 per genre).
  - Clips from the GTZAN dataset.
- https://www.kaggle.com/datasets/solomonameh/spotify-music-dataset
  - Spotify proprietary spotify features for ~1700 pop songs

https://www.kaggle.com/datasets/anandshaw2001/netflix-movies-and-tv-shows

https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data

By February 14 each team will submit a short project proposal. The project proposal shall include the following information:

Names of all team members.

Dataset information: describe the dataset/datasets you plan on using for the project.

Questions to study. Outline the analytical questions you want to work on as part of the project.

The proposal shall be word-processed and submitted as a PDF document. It should be relatively short.