Potential Ideas

<https://data-flair.training/blogs/data-science-project-ideas/>

Music datasets

<https://lionbridge.ai/datasets/12-best-audio-datasets-for-machine-learning/>

Prior GA Projects

<https://gallery.generalassemb.ly/DSI?metro=>

Ideas from classmates

* -Patrick’s suggestion: allow a model to use image data to predict what a person is interested in (e.g. if a person is in a group of people, do they often purchase certain products over others)
* If you are using Kaggle data, you *must also* add your own data or ensure that you are cleaning some data

**Questions for Greg**

* Business impact, creativity, and how good it looks visually – most important features
* Try to avoid a recommendation system
* Take a video of you presenting the Powerpoint and put it up on Youtube
* Explore public datasets around Hulu -

Entertainment industry:

Links:

Movie recommendation idea:

<https://www.kaggle.com/rounakbanik/movie-recommender-systems>

<https://data-flair.training/blogs/data-science-r-movie-recommendation/>

<https://www.lafabbricadellarealta.com/open-data-entertainment/> > use this link to search for movie datasets

Prompt Ideas:

* Predict genre of a movie given script
* Predict movie rating given various features
* Predict gross revenue of movie given various features
  + Link to potential dataset: <https://www.kaggle.com/danielgrijalvas/movies>

-customer segmentation: identify people to put into groups (clustering); after this, look at what the customers have in common

-use clustering as initial step to identify groups and then feeding this into a recommendation system; *combination of clustering (grouping) AND THEN use this to a recommendation system*

-will need to find user data OR it takes in characteristics of movies and find relationship between features; once relationships are established, you can take in user input and provide recommendations based on these relationships

~~Potential Movie Datasets~~

* [~~https://www.imdb.com/interfaces/~~](https://www.imdb.com/interfaces/) ~~> could be promising since I could join the datasets on the~~ **~~titleId~~** ~~to allow myself more features~~
* [~~https://grouplens.org/datasets/movielens/~~](https://grouplens.org/datasets/movielens/) ~~> very good dataset but I had to send a request form….(I sent request form for MovieLens 25M Dataset)~~
* [~~https://www.kaggle.com/tmdb/tmdb-movie-metadata?select=tmdb\_5000\_movies.csv~~](https://www.kaggle.com/tmdb/tmdb-movie-metadata?select=tmdb_5000_movies.csv)
  + ~~Good basic data on budget, general review, description of movie, genre~~
  + ~~No info on user data~~

Dataset to use

* <https://data.world/robertjoellewis/film-subtitles> > Linguistic data that has a lot of great info > would need to take the project in a different direction though
  + Link for documentation of this dataset: <https://www.tandfonline.com/doi/abs/10.1080/10510974.2017.1340903?journalCode=rcst20>
  + Word count dictionary: <https://moralfoundations.org/>
    - <https://moralfoundations.org/wp-content/uploads/files/downloads/moral%20foundations%20dictionary.dic>
  + Ex problem statement: given certain ratios of text, we can predict how successful a certain genre of movie will be

Other ideas that could even be added on to it if time permits

-image classification on movie poster; or NLP on plot summary of movies;