**Final Project Proposal**

**Understanding Film Rating Data**

**Team members: Eduarda Espindola, Payman Roghani, Mark Yong**

W200 Python Fundamentals for Data Science, Summer 2018

UC Berkeley MIDS

**Objective:**

With the abundance of channels available to watch movies, it is challenging for consumers to find movies that match their preferences and tastes. Some platforms (e.g. Netflix) provide recommendations based on users’ ratings applied to films or shows they watch, as well as the content they have previously watched. The purpose of this project is to better understand how users apply ratings and tags to movies, and find potential correlations between ratings and other parameters.

**Key areas of focus:**

* Correlation between ratings and movie genre
* Correlation between ratings and movie release date
* Distribution of ratings (i.e. are most ratings either very positive or negative?)
* Average rating per user
* Correlation between user rating and user tagging
* Correlation between number of ratings for a movie with the average rating for that movie (i.e. does a great number of user ratings for a movie mean higher average rating for that movie?)
* Comparison of ratings on the 3 platforms we will evaluate (MovieLens, TMDB, IMDB). If different, what are the possible reasons? For instance, are different platforms biased towards specific genres, directors, etc.?

**Datasets:**

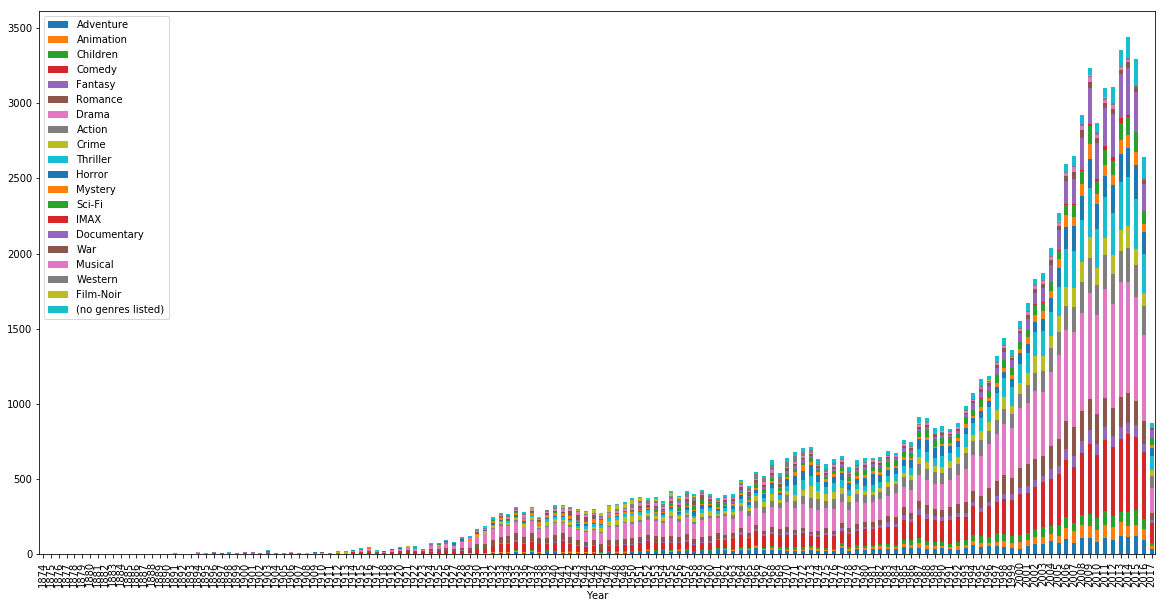
* Primary source:
  + MovieLens
    - Dataset: MovieLens 20M Dataset (includes 20 million ratings and 465,000 tag applications applied to 27,000 movies by 138,000 users)
    - Link: <https://grouplens.org/datasets/movielens/20m/>
    - Key columns in the datasets: movie title, release date, genre, tags, ratings, user ID, TMDB ID, IMDB ID
* Other sources:
  + TMDB (The Movie Database):
    - Key data for our project: ratings, cast and crew, plot keywords, similar movies, reviews
    - Link to API description: <https://www.themoviedb.org/documentation/api?language=en>
  + IMDB (Internet Movie Database)
    - Key data for our project: ratings, runtime minutes, movie characters
    - Link: <https://datasets.imdbws.com/>
  + Kaggle The Movies Dataset
    - Already have some links between the movielens database and TMDB
    - Link: <https://www.kaggle.com/rounakbanik/the-movies-dataset/home>

**Approach and charts (to be provided by Eduarda)**

We will analyse rating trends between users (and/or groups of users - the clustering factor yet to be defined) in order to assess particular trends of ratings per user, which could lead to an effective prediction system: if user X gave those ratings to those movies, to movie Y, user X will most likely give a rating between A and B to movie Y. That may lead to a more effective suggestion engine for movies.

Besides the suggestion engine, this could lead to a better understanding of trends in successful movies for the industry: some genres with particular storyline elements and set of actors and directors tend to be more successful than others.

By that approach, let us begin by an initial analysis of movie genres over the years:



By that chart we are able to see how the representativity of drama movies have increased over the years.

And an average grade per genre over the years: