**Project Movie Prophet - Feature Engineering Details**

**Background**:

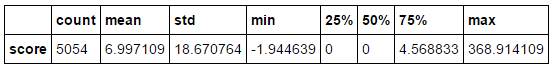
Data scraped from 3 sources - Box Office Mojo, Wikipedia and Open Movie Database.

**Features Created:**

1. **Players**

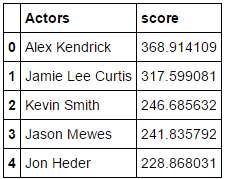
Players includes Actors, Cinematographers, Composers, Directors, Producers and Writers. The methodology for computing scores for each of the players is the same. Below, we are only considering Actors.

* 1. **Methodology**
     1. Considered 16828 movies from Box Office Mojo
     2. Filtered only the movies having information on budget and total gross. Total movies remaining is 3066.
     3. For the 3066 movies, calculated the ‘movie success’ using the formula *(movie\_total\_gross - movie\_budget) / movie\_budget*. Sign of the result indicates positive/negative revenue and magnitude indicates amount of revenue.
     4. For each unique actor, the ‘actor\_success’ score was calculated by adding the movie\_success score for all the movies the actor featured in
     5. Extracted all unique actors in all movies. Total unique actors is 5054.
  2. **Improvements Planned**
     1. Weigh the score based on whether an actor had a minor role in the movie. Actors who had minor role get lower score.
     2. Weigh the score based on the date the movie was released. Recent movies get higher score.
     3. Take into consideration awards
     4. Take into consideration IMDB rating
     5. Feature is biased towards actors who appear in more movies so that needs to be normalized.
  3. **Results**
     1. **Summary**: A total of 2917 Actors out of 5054 have a score of zero meaning their movies did not have information for budget and total gross
     2. **Statistics for score**:

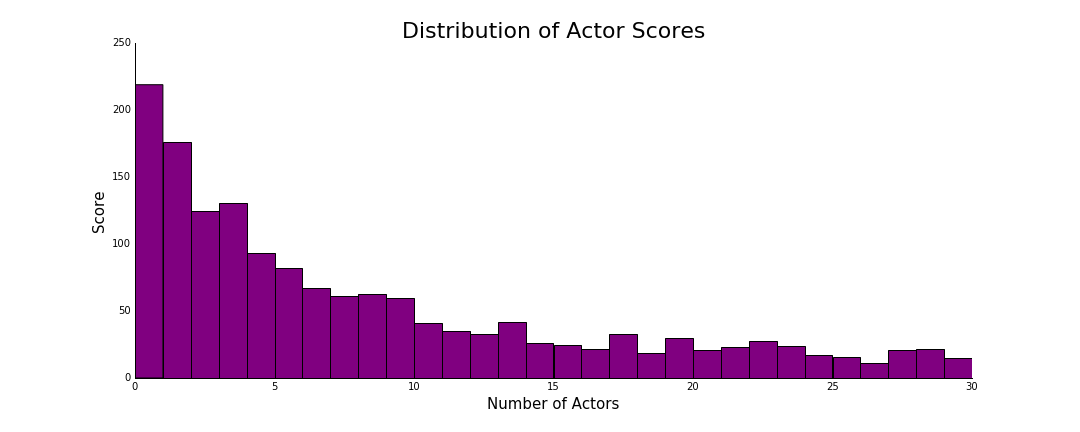


.

* + 1. ***Top Actors*** :



* + 1. ***Distribution***: For the distribution below, actors with score of zero have been ignored.



1. **Response Variable: Revenue**

Filtered the data for movies containing only budget and gross information. This resulted in 6754 unique movies. For each of the movies, the revenue was calculated as:

Revenue = (gross-budget)/budget

Note:

1. Budget and Gross information was present only in wikipedia and box office mojo. Thus the movies were filtered from box office mojo and wiki
2. For a particular movie, if budget and gross information are present both in bom and wiki, the data was fetched from bom.
3. Box\_domestic\_gross was the bom feature used for gross.
4. **Features related to release date:**
5. Month of the year
6. Releasing week (from 1 to 53)
7. Release quarter (1,2,3,4)
8. Release day (no:of days since Jan 1 of the release year)

Note:

1. Release date information was present in omdb, bom and wiki. Preference given to bom>omdb> wiki.
2. The release date field from wiki was messy. Conversion to datetime led to loss of information.
3. **Holiday season**
4. Based on the release date, created a binary feature holiday\_season indicating whether the release of the movie was during holidays (1) or not (0).
5. Considered all the US Federal holidays from the year starting 1900
6. Movies that released one week prior to the US Federal holidays also labeled as released during holiday\_season