Marketing Analytics
Final Project Checkpoint 2
Jacob Dineen, Ruben M Suzara, Micaela Geiman, Samuel Harvey, Fiona Erickson
Due: 5/3/2018

Project Plan and Research Design: Group3

About our data:
Our group has decided on an extension of the 'Movie Information' dataset provided to us in this class, instead relying on a popular dataset from Kaggle, found here. The dataset was initially scraped from IMDB, but due to a takedown request is now sourced from The Movie Database (TMDb). The Kaggle dataset is comprehensive, containing 5000 films released from the early 1900's through 2016, and containing 20+ features, including, but not limited to, Box Office Gross, Budget, Top 3 Billing Actors, Director, Average Voting Score and Total Number of Votes.

Data cleaning/Preprocessing will utilize existing scripts from Kaggle kernels to ease the burden of text parsing and concatenation of multiple datasets. Most text columns are scraped in JSON format. There are scripts available on Kaggle to clean the data and revert it to its original, pre-scraped format (dictionaries for text features). Popular way to clean this data found here.

Research Plan:
Each member of the group will be working on a separate subproject which will be consolidated at the end of the term and presented by each member. Analysis/Insights/Documentation on each 'subproject' will be noted here. Subprojects will contain elements of machine learning (descriptive/predictive/prescriptive analytics), statistical analysis, marketing recommendations and adhoc visual analysis pertaining to each subproject goal. Marketing recommendations/solutions will stem from said analysis in a variety of different forms. Some potential ideas:

1. Content Based Recommendation Engine based on plot keywords, actors, directors, popularity/voting score. Will use popular python NLP libraries for stemming/tokenization of text. Likely to utilize distance based algorithm (nearest neighbors) for matching.
2. Run regression analysis to see what factors are significant in predicting profitability or box office gross- analyze omitted variable bias. Perform logit to emphasize feature importance in the model.
3. Create visualizations on movies released by year, trends in popularity across genre and time. Making different charts and graphs to showcase our data.
4. Cluster analysis segmented movies into different buckets.
5. Binning popularity into two classes and doing a straightforward class problem w/ logits/random forests/neural nets. Potentially doing logistic regression for isprofitable/isnotprofitable
6. Correlation analysis to understand which features share relationships, and exposing multicollinearity - Feature engineering on regressors to eliminate multicollinearity.
7. Standard time series analysis to understand and visualize the market over time.
8. Chi square hypothesis testing to signify observed values against expected values and determine causal relationships.

Features in our dataset: Budget, Genres, Keywords, Language, PlotOverview, Popularity, Production Companies, Production Countries, Release Date, Duration, Box Office Gross, Spoken Languages, Title, Average Votes, Count of Votes, Country, Director Name, Top 3 billed actor names

From our first checkpoint:
If our models are able to generalize to unseen data, we'd have the predictive capabilities to do things like:

- predict box office gross based on seasonality/director/top billed actors
- predict average voting scores based on plot keywords
    - Understand word density in movie taglines, and how they correlate to dependent variables. This is useful in marketing campaigns for prereleases.
- Recommend movies based on similar characteristics, such as plot keywords, taglines, release dates, genres.
- Understand the relationship between user ratings and metrics of success like box office gross and profitability - Can bad word of mouth kill a movie release?

Some idiosyncrasies of our data to note:

- Monetary values are not adjusted for inflation. May be worthwhile to scale them at a hierarchical level (title_year).
- There are a couple of known outliers that we will need to remove.
- No user data available, so Collaborative Filtering is not an option for reco engines. Would have to be content based (plot keywords).