Marketing Analytics
Final Project Checkpoint 1
Jacob Dineen, Ruben M Suzara, Micaela Geiman, Samuel Harvey, Fiona Erickson
Due: 4/21/18

Project Idea and Team Composition: Group3

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.

This dataset is of high interest due to the variety of machine learning applications which can be applied to it, and the relationship between some of these solutions and the world of marketing. In particular, we'd like to explore some of the following aspects of this dataset:

1. K means Clustering based on various features.
2. Regression on rating/popularity against budgets/actors/studios/release dates
3. Binning popularity into two classes and doing a straightforward class problem w/ logits/random forests/neural nets
4. text mining taglines w/ term frequency matrices and maybe predicting rating or profitability.
5. Visualizations on movies released by year, trends in popularity across genre and time, etc..
6. Potentially building a lightweight recommendation engine based on plot keywords. This is more advanced, but is done here.

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).

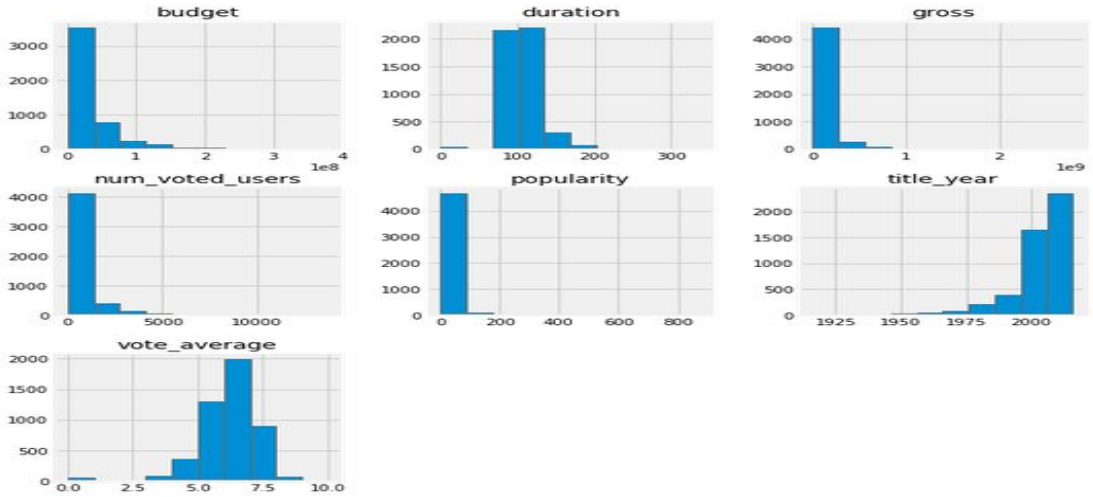
Most text columns are scraped in JSON format. There are scripts available on Kaggle to clean the data and revert it to its original, pre-scraped format (dictionaries for text features). Popular way to clean this data found here. Being an open source project, there are many solutions already available to help us if we get stuck, but we will try to deviate from verbatim code/analysis to produce original perspectives.

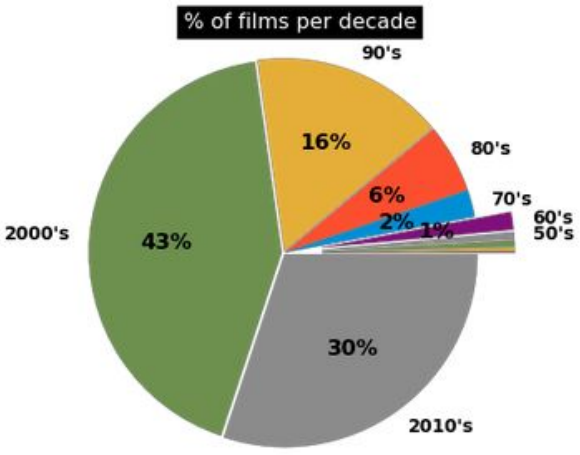
Tools needed: Python, Weka, R, Excel, XLStat

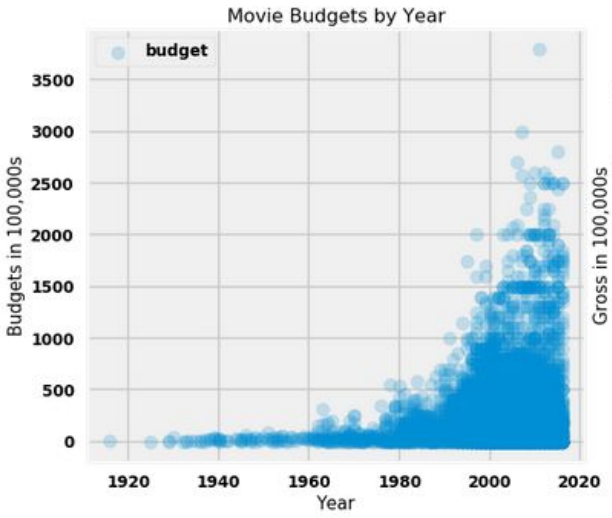
Machine Learning Techniques: Logit Models, Neural Nets, Natural Language Processing/Text Mining, Clustering, Association Rules/Recommendation Engines

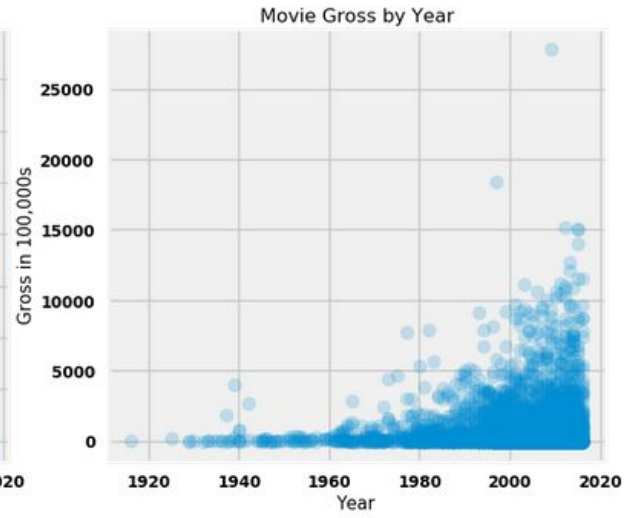
Resources: StackOverflow, R/Py documentation, Course Materials

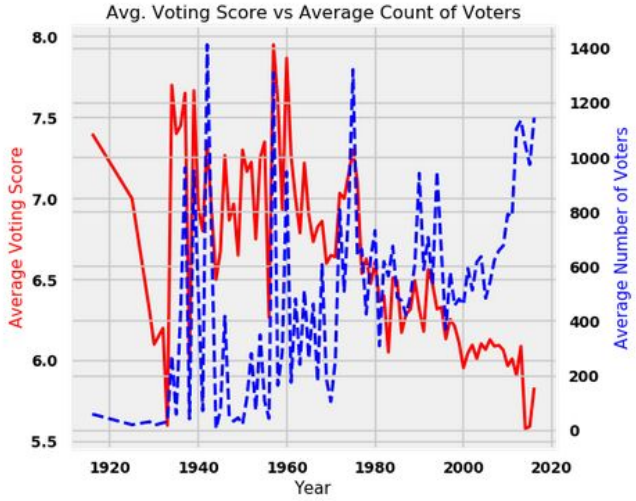
Initial Data Visualization (base py/matplotlib/seaborn):

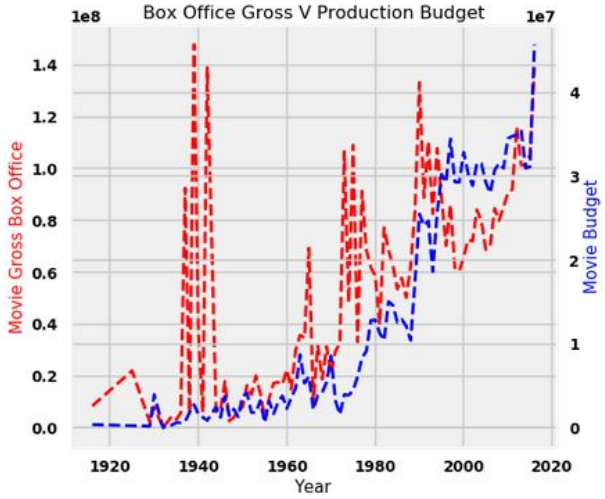





Movie Budgets by Year
budget
Budgets in 100,000s
Year


Movie Gross by Year
Gross in 100,000s
Year


Avg. Voting Score vs Average Count of Voters
Average Voting Score
Average Number of Voters
Year


Box Office Gross V Production Budget
1e8
1e7
Movie Gross Box Office
Movie Budget
Year