Fall 2016 OMS DAVA Project - Part I

Submission Deadline: 2016-10-10 (see T-Square for specific time)

The file `movies_merged`[1] contains a dataframe with the same name that has 40K rows and 39 columns. Each row represents a movie title and each column represents a descriptor such as `Title`, `Actors`, and `Budget`. I collected the data by querying IMDb's API (see `omdbapi.com`) and joining it with a separate dataset of movie budgets and gross earnings (unknown to you). The join key was the movie title. This data is available for personal use, but IMDb's terms of service do not allow it to be used for commercial purposes or for creating a competing repository.

The project's theme is to investigate the relationship between the movie descriptors and the box office success of movies, as represented by the variable `Gross`. This task is extremely important as it can help a studio decide which titles to fund for production, how much to bid on produced movies, when to release a title, how much to invest in marketing and PR, etc. This information is most useful before a title is released, but it is still very valuable after the movie is already released to the public (for example it can affect additional marketing spend or how much a studio should negotiate with on-demand streaming companies for "second window" streaming rights).

Part I of the project involves getting the data ready for analysis and modeling and doing some preliminary investigations. It is 50% of the total project grade. Part II will involve modeling and predictions and will be released at a later date. Please complete the assignments below and submit a PDF file containing answers as well as a separate .R file containing the code that you used with adequate documentation showing which code was used for which of the 8 assignments below. The code can assume that the data file is already loaded but should have all `library` commands at the beginning and be self contained.

The assignments below have equal weight and are sequential i.e., do step 2 after you processed the data as described in step 1. It is OK to handle missing values below by omission, but please omit as little as possible. It is worthwhile to invest in reusable and clear code as you may need to use it or modify it in part II of the project.

Assignments:

1. The variable `Type` captures whether the row is a movie, a TV series, or a game. Remove all rows that do not correspond to movies. How many rows did you remove?
2. The variable `Runtime` represents the length of the title as a string. Write R code to convert it to a numeric value (in minutes) and replace `Runtime` with the new numeric column. Investigate and describe the distribution of that value and comment on how it changes over years (variable `Year`) and how it changes in relation to the budget (variable `Budget`).
3. The column `Genre` represents a list of genres associated with the movie in a string format. Write code to parse each text string into a binary vector with 1s representing the presence of a genre

---

[1] Available at `https://s3.amazonaws.com/content.udacity-data.com/courses/gt-cs6262/project/movies_merged`.

and 0s the absence and add it to the dataframe as additional columns. For example, if there are a total of 3 genres: Drama, Comedy, and Action a movie that is both Action and Comedy should be represented by a binary vector (0, 1, 1). Note that you need to first compile a dictionary of all possible genres and then figure out which movie has which genres (you can use the R `tm` package to create the dictionary). Graph and describe the relative proportions of titles having the top 10 genres and examine how the distribution of gross revenue (variable `Gross`) changes across genres.

4. The dataframe was put together by merging two different sources of data and it is possible that the merging process was inaccurate in some cases (the merge was done based on movie title, but there are cases of different movies with the same title). The first source's release time was represented by the column `Year` (numeric representation of the year) and the second by the column `Release` (string representation of release date). Find and remove all rows where you suspect a merge error occurred based on a mismatch between these two variables. To make sure subsequent analysis and modeling work well, avoid removing more than 10% of the rows that have a present `Gross` variable. What is your precise removal logic and how many rows did you end up removing?
5. An important question is when to release a movie. Investigate the relationship between release date and gross revenue and comment on what times of year are most high revenue movies released in. Does your answer changes for different genres? Based on the data, can you formulate a genre-based recommendation for release date that is likely to increase the title's revenue? If you have a recommendation motivate it with the appropriate disclaimers, or otherwise explain why you are unable to produce a recommendation.
6. There are several variables that describe ratings including IMDb ratings (`imdbRating` represents average user ratings and `imdbVotes` represents the number of user ratings) and multiple Rotten Tomatoes ratings (represented by several variables pre-fixed by `tomato`). Read up on such ratings on the web (for example `rottentomatoes.com/about` and `http://www.imdb.com/help/show_leaf?votestopfaq`) and investigate the pairwise relationships between these different descriptors using graphs. Comment on similarities and differences between the user ratings of IMDb and the critics ratings of Rotten Tomatoes. Comment on the relationships between these variables and the gross revenue. Which of these ratings are the most highly correlated with gross revenue (use the R function `cor` and remove rows with missing values)?
7. The variable `Awards` describes nominations and awards in text format. Convert it to a three dimensional binary vector whose first component represents no nomination or awards, the second component represents some nominations/awards, and the third component represents many nominations or awards. The relationship between the second and the third categories should be close to 5:1 (not precisely - this is a broad guideline to help you avoid creating a third category that is useless due to being extremely small and to encourage consistency). How did you construct your conversion mechanism? How does the gross revenue distribution changes across these three categories.
8. Come up with two new insights (backed up by the data and graphs) that are expected, and one new insight (backed up by data and graphs) that is unexpected at first glance and do your best to motivate it. By "new" here I mean insights that are not an immediate consequence of one of the above assignments.