Data Cleaning the Movies Datasets

I obtained two datasets containing both general movie data and data indicating whether films were nominated for academy awards. The general movie dataset contains imdb data obtained from Kaggle at https://www.kaggle.com/antoniobap/imdb-v3/data. The academy award data was obtained from the University of Waterloo at https://cs.uwaterloo.ca/~s255khan/oscars.html.

Before loading the datasets, I loaded the dplyr, tidyr and dummies libraries.

Preliminary cleaning procedures

Intending to merge these two datasets into one, I began by changing the names of the columns in the imdb dataset (hereinafter, imdb) to match those in the academy awards dataset (hereinafter, acad_awd). I renamed movie_title, title_year and imdb_score to be name, year and rating, respectively.

I then trimmed all leading and trailing whitespace from the names in both datasets and removed trailing question marks from the names in the imdb set. I also changed the names in both datasets to uppercase for consistency. I changed one title in imdb, Birdman, which used the film's extended title to match the shorter title used in acad_awd. There were two dashes in the nominations column in the acad_awd set that I changed to 1's, signifying that those films had at least one nomination for Best Picture.

The na's in certain key columns in imdb, budget, gross and duration, had values imputed to them based on the means of their respective columns as a whole. (The mean for duration was rounded to ensure an integer value.)

A number of duplicated names were observed in the imdb set as well, and these were removed using imdb <- distinct(imdb, name, .keep_all = TRUE). I then subsetted the imdb data for durations greater than 60 in order to eliminate the television shows that were present in the data, and then removed blank names.

The Merge Process

The merge process proved to be fairly complex owing to the different columns in each dataset. In order to prevent loss of data from the acad_awd set, the merge was accomplished in two distinct stages.

First, a merge_acad_awd set was created, dropping all columns except the name and nominations columns. Following this, I created a merged dataset called movies by performing a full join of merge_acad_awd and imdb by name (movies <- full_join(merge_acad_awd, imdb, by = "name")). This merge was effective except for those movies that were in the acad_awd set, but not in the imdb set. All data except for their names and nominations were effectively lost in the merge. As their were 17 such films (out of a total of 86 acad_awd films), this proved to be an unacceptable loss.

Accordingly, a second separate merge of the 17 films into movies was necessary in order to at least preserve duration, ratings and genre-related data. In order to make the genre-related data in acad_awd for these films compatible with the movies set, I created binaries from the genre1 and genre2 columns using the dummies library. These binaries had to first be extracted into 2 separate dataframes, binary1 and binary2, in order to prevent overwriting in situations where a film was both a comedy and a drama, for example. The two binary sets were then bound together using cbind and the column of blanks, V1, was removed. The combination of the two sets in this manner created duplicate column names such that, for example, Biography could be added to Biography.1 (as in binaries$Biography <- binaries$Biography + binaries$Biography.1) so that any ones would cancel out any zeroes in order to allow for the presence of two genres simultaneously. The duplicate columns were then removed. I then attached this final set of binaries to acad_awd using cbind and called the resulting dataset merge2_acad_awd. I then dropped all extraneous columns from merge2 (genre1, genre2, release, metacritic and synopsis) and limited this dataset to only those films that did not appear in imdb (merge2_acad_awd <- merge2_acad_awd[which(merge2_acad_awd$name %!in% imdb$name),]). I then performed the second merge by doing a full join of merge2 with movies, initially into a test dataset called movies2. After examining movies2, I overwrote movies with movies2. This final merge proved to be successful as the maximum amount of data appeared to be preserved.

Final preparations

Following the final merge, I created two subset dataframes for training and testing purposes. The first set, to be used as a training set, consisted of all movies between 1927 and 2014 inclusive (the years common to both the original datasets) and ultimately called train_movies. Any films in this set having NA's in the nominations column were changed to 0 to signify that they were not nominated. The second set consisted only of newer movies (test_movies), and naturally had no nominations data. This set could be used for testing and verification after analysis and modeling of the training set. Both sets then required a small amount of additional cleaning to remove additional NA's and irrelevant columns (due to lack of relevancy to older films, such as Facebook likes, for example), after which they were ready for analysis.