Report on Movies Dataset

The dataset contains the information of around 8,000 movies from the top 10,000 movies by revenue according to The Movie Database. Part of the data was gathered from The Movie Database API (Title, Budget, Revenue, Data, Runtime) which created the initial 10k movies, the other part was then gathered from OMDB API filling the rest of the features, at that point we had the original unfiltered dataset - can be seen by opining the “Raw Movies.csv” file – then we looked at our dataset and examined the problems with it, and we found the following:

1. The Year column received from the OMDB API was missing a lot of values, where the one got from The Movie Database was surprisingly complete, so we dropped the year column and extracted only the year from the date column, then renamed it into “Year of Release” to better reflect the actual content of the column.
2. The Language column was missing data for some movies. Since all the movies considered were from Hollywood, furthermore, the column meaning is: in what languages was this movie available? We decided to fill the missing values with English, based on the logical idea that all movies released in Hollywood were released in English at the very lease.
3. The Country column was treated in a similar way to Language column
4. The Awards column describes what awards was this movie nominated to/awarded. So, for movies with this value missing we assigned the value None
5. The IMDB Rating & Votes columns: For movies with this column missing we assigned the value zero, since this movie was not rated on IMDB and no actual movie has a value zero, making it not meaningful in our dataset.
6. The Actors column: This column had the names of the most famous actors in that film. This was alright, but was going to be a problem in the future (in the next part of the project), so, since we don’t actually care about the names of the actors (not the aim of this project) and we only care whether this movie had any top 100 actors, we decided to only count the number of actors in the top 100 actors of all time (based on IMDB) and put the count instead of the names of the actors.
7. The Directors & Writers columns: Those columns were initial going to be treated similarly to the actors column, but after we saw the result, we found that extremely low number of movies had their writers or directors in the top 100. Furthermore, no concreate list of top 100 directors or writers by success of their movies or popular votes could be found. Hence, we decided to drop those columns.
8. The Production column: Some movies had this value missing, and since their number was not huge, we decided to give them the value (“Other”) for this column to reflect that it’s not one of the recurring production companies that repeated in most of the other movies.
9. The Runtime column: For movies missing this value we decided to assign the mean of run time (around 108 minutes interestingly).
10. The Rated column: This column describes the Motion Picture Association of America film rating for this movie. A small section of movies collected from the Movie Database had other ratings considered like “Passed” or “Approved” so we decided to change this to Unrated to make it uniform in the same rating system.

These changes lead to the filtered dataset that we will consider moving forward in this project. This dataset is called “Filtered Movies.csv”.

Note that we included both datasets to give you the ability to see the difference between the raw or original dataset, and out filtered dataset.

Ahmad Nawar Droubi – 1600000631

Aydın Davutoğlu - 1600000117