IMDB and Social Media Project Proposal

Group 4

9 March 2019

Note:
- No need to specify concrete statistical methods (yet!)
- Not more than 1-2 pages
- One proposal per group

1. Background to data set – why is this an interesting data set?

The dataset proposed for this analysis is the *IMDB 5000 Movie Dataset* (Sun, 2016) that was obtained online from the data.world repository. The creator, Chuan Sun, was able to scrap metadata from a combination of three sources: www.the-numbers.com, IMDB.com, and the Python library, "scrapy". All together, information regarding 28 variables was obtained for 5043 movies. These observations included information from some 4906 posters, spanning across 100 years in 66 different countries, with 2399 unique director names, and thousands of actors and actresses. Social media content, specifically related to the Facebook likes of the movie and its cast were also included. This dataset offers an interesting insight into possible relationships between movie charateristics such as where, when and by who it was made, and how the movie is perceived by the public on social media.

I don't know if this is actually necessary:
Of the 28 variables, 14 were numeric while the other 14 were factor variables. An initial inspection of the data revealed that some variables were unnecessary, such as the IMDB link to the movie, while others were too complex or unique to model, such as plot keywords. In total, there were 2698 missing values across the entire dataset. However, with careful data cleaning techniques this should be simple to remedy without too much loss of information.

Possibly more important to talk about why it is actually interesting?

2. Why is it a multivariate data set?

Multivariate data is the data in which analyses are based on more than two variables per observation. In the case of the IMDB movie dataset, there are clearly multiple variables to work with, hence, fulfilling the basic criteria to be multivariate. However, the dataset must also be multivariate in the sense that the relationships between variables are not univariate, i.e. it would not be appropriate to model with a linear model. With the dataset in question we have reason to believe that the relationship between certain variables is not univariate. For example, the gross income may be modelled using multiple other variables, however, other variables in the dataset may in turn be affected by the gross income, perhaps the IMBD score. This suggests that it may be inappropriate to model any particular variable in the dataset with a linear or generalised linear model. Instead, multivariate methods such as PCA or a clustering algorithm may be more appropriate.

It is with the above points in mind that we believe there may be an underlying structure in the data, whether it be a correlation between multiple variables or an unexplained trend. There have been many cases when movie critics have given poor reviews and yet the movie has gone on to be a box office success. This gives us more reason to believe that there is additional structure in the dataset which hopefully can be explored using the social media data. By combining the analysis of the continuous variables with the factor variables, we hope to uncover any hidden structure.

3. General aim of the future multivariate analysis (not set in stone!)

Because of the popularity of social platforms, the social media campaign now becomes one of the most important marketing tactics for movie industry. According to previous researches, such as Pardo's paper (2013), social media have great impact on movie qualitatively. However, in this study, we aim to investigate the quantitative relationship between social media and movie performance such as box office and movie ratings.

Possible Edit: Due to the increasing popularity of social platforms, the social media campaign has now become an important marketing tactic in the movie industry. Previous research, such as Pardo's paper (2013), found that social media has a great impact on movie qualitatively. In this study, we aim to investigate the quantitative relationship between social media and movie performance such as box office and movie ratings.

Reference List

Pardo, A., 2013. Digital Hollywood: How Internet and Social media are changing the movie business. In Handbook of Social Media Management (pp. 327-347). Springer, Berlin, Heidelberg.

Sun, C., 2016. IMDB 5000 Movie Dataset. data.world. e675d8a8. Available at: https://data.world/popculture/imdb-5000-movie-dataset