Project Proposal: IMDB Movie Anaylsis

Group 4

8 March 2019

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 and crew 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.

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 5000 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 there is reason to believe that the relationship between certain variables is not univariate. For example, the gross income may be modelled using multiple other variables, which 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 clustering algorithms 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!)

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.

Our second general aim would be to investigate the relationship among film box office grosses, budget and the county of origin. Hollywood's stereotypes seemingly dominate in reality, while most films around the world are produced from English-speaking countries. In the wave of globalisation and with the popularity of social media nowadays, we wish to question if cultural diversity in film industry still exists as in other forms of arts. Regarding inter-country, or in a broader sense, inter-continent cultural difference, we think that a principle component or cluster analysis would be a good starting point. The characteristics of the mainstream film in various countries should reflect the colour of their culture. We also would like to have a

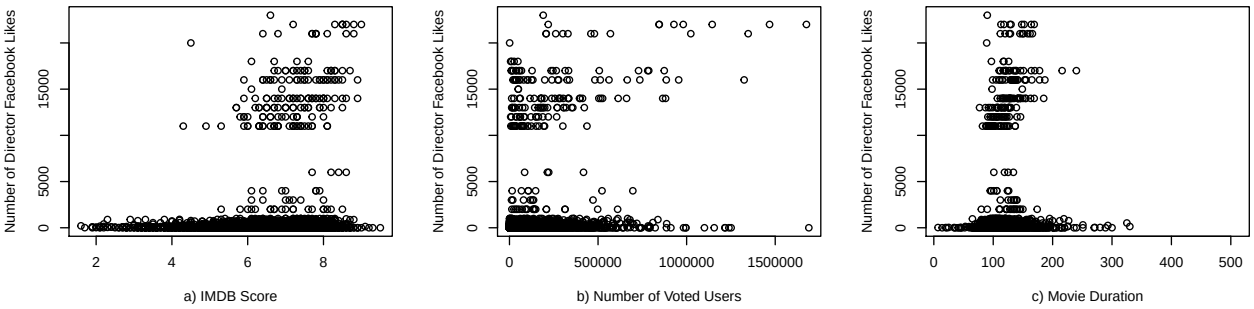

Figure 1: Paired Scatterplots of The Number of Director Facebook Likes and Various Movie Characteristics

glimpse into the niche markets for film industry. These films of interest could be non-mainstream, low-budget or independent film. We would examine if these films could satisfy specific market needs.

Figure 1 displays three scatterplots that show the relationship between the Facebook likes received by the director and IMDB score, the number of users that voted and the duration of the movie. The plots indicate that there may be distinct groups relating to a lower and higher number of director Facebook likes. However, it is not clear what may be causing these groups. Through further analysis of these relationships, and other similar relationshps we aim to uncover hidden multivariate trends between social media and movie characteristics.

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.