Project Proposal: IMDb Movie Anaylsis

Group 4: Brooke Franklin, Carson McKee, Manho Suen, Yangzhu Chen, Shuang Qiu

8 March 2019

Background to the *IMDB 5000 Movie* Dataset

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 scrape 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 Internet Movie Database (IMDb) rated movies. These observations included information from some 4906 posters, spanning across 100 years in 66 different countries, with 2399 unique director names as well as those for 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 characteristics such as where, when and by who it was made, and how the movie is perceived by the public on social media.

Multivariate Data

In order to be considered multivariate, a dataset needs to meet two criteria. Firstly, analyses are based on more than two variables per observation and secondly, the relationships between variables are not univariate, but rather there are a combination of relationships between variables.

In the case of the *IMDB 5000 Movie* dataset, there are clearly multiple variables to work with, hence, fulfilling the first basic criteria to be multivariate. More importantly, there is reason to believe that the relationships between certain variables are more complex than being simply univariate and there may be an underlying structure in the data, whether it be a correlation between multiple variables or an unexplained trend. This suggests that it may be inappropriate to model any particular response type variable in the dataset with typical modelling methods, such as a linear or generalised linear model. Rather, we would prefer to uncover these interconnected relationships using more appropriate multivariate methods such as Principal Component Analysis (PCA) or clustering algorithms.

General Aims of the Analysis

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 movies, qualitatively. In this study, we aim to investigate the quantitative relationship between social media and movie performance such as box office and movie ratings. 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 related covariates. Thus, by combining the analysis of the continuous variables with the factor variables, we hope to uncover any such hidden structures.

Further to this, we aim to explore the relationships among box office grosses, budget and the country of origin. An exploartion into inter-country, or in a broader sense, inter-continent differences between movie performace and characteristics, such as non-mainstream films, low-budget or independent films, and their corresponding public perception on social media could expose an interesting viewpoint on the diversity of IMDb rated movies.

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

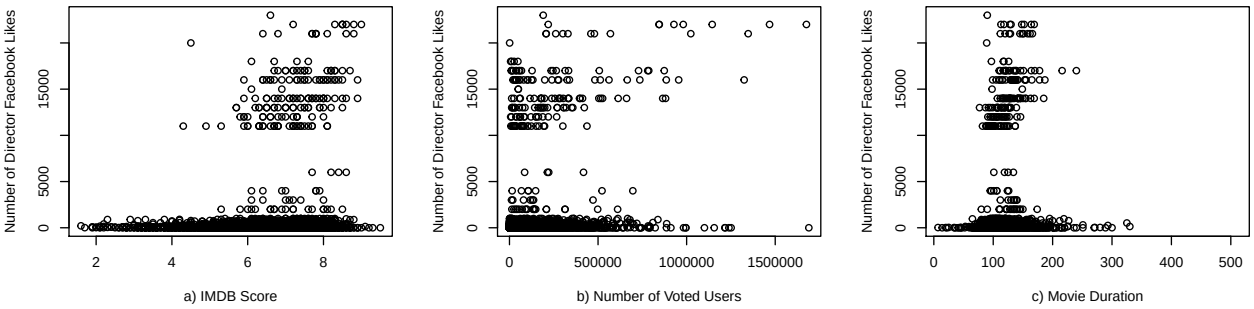

Figure 1: Paired Scatterplots of The Number of Director Facebook Likes and Various Movie Characteristics

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, and other similar relationshps, we aim to uncover hidden multivariate trends within the *IMDB 5000 Movie* dataset.

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.