Why is the dataset multivariate?

Multivariate data is the data in which analysis 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 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.