# *K*-means clustering of in-theater movie competition to predict box office sales

Lance Einfeld
Mentored by Richard Latham
Booz | Allen | Hamilton

## Introduction

The movie industry is a multibillion-dollar business with individual movies costing tens of millions of dollars to produce and market. Movies generally last in theaters for ten weeks, and as a result, new movies are frequently introduced to maintain business (Jedidi, 1998). Due to their high production cost, movies unavoidably pose a risk of loss along with their chance of profit. The ability to reduce the risk of loss while also maximizing sales of a movie would prove to be invaluable to studios in the movie industry. The goal of this study was to model the box office success of a movie in the United States of America based on the competition it faced in theaters. It aimed to find a correlation between factors that affect a movie's competition and box office sales through the use of *K*-means clustering and multiple linear regression (MLR). This research differed from past research aiming to predict sales by focusing solely on the effects of various competition factors; this research focused on the environment a movie faced more than the movie itself. This study also incorporated online reviews, which have increased in significance along with the increase in internet use.

## Materials and Methods

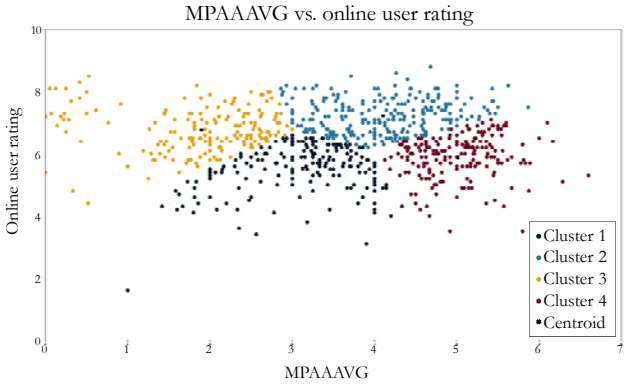
Movie data was collected for all movies that reached the top ten on the weekly movie charts for all weeks in the five-year time period ranging from January 1st, 2010, to December 31st, 2015 (808 movies). Data collection was performed using a web-scraping script written in the Python programming language with the BeautifulSoup package that enables the extraction of data from html files. The web-scraping script collected data from two online movie box office databases. From the first database, the script collected the movies' titles, total grosses, genre, MPAA rating, production budget, release date, and weekly grosses. From the second database, the script collected the movies' online critic review ratings and online user review ratings.

A second Python script was written to calculate metrics of competition for each movie. GENRETOT was defined as the total number of movies in the top ten with the same genre for all weeks of showing. MPAATOT was defined the same as GENRETOT except with MPAA ratings. MPAAAVG and GENREAVG were defined as MPAATOT and GENRETOT divided by the number of showing weeks, respectively. SLEEPERSCALE was defined as the percent of the total gross of a movie from its opening week alone.

## Materials and Methods (cont.)

SEASON was defined as the number of days a movie was released from November 20th of the prior year. The script also filtered out movies with missing data (671 movies remaining). A third script in Python was coded for data visualization, and for performing *K*-means clustering and MLR. The results of the *K*-means clustering were verified with Wolfram Mathematica®, and the results of the MLR were verified with R®.

## Results




Graph 1 (above): *K*-means clustering performed on the sample with MPAAAVG and online user rating as the dimensions. The centroids for cluster one, cluster two, cluster three, and cluster four are (3.126, 5.519), (4.103, 7.219), (1.896, 6.770), and (5.013, 5.885), respectively. The clusters lack distinct separations from one another.

| | Unclustered | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|---|---|---|---|---|---|
| Intercept | $3.59 \times 10^{-16}$ | 0.000244 | 0.00162 | 0.00259 | 0.000638 |
| GENRETOT | $4.38 \times 10^{-7}$ | 0.649777 | 0.03156 | $1.43 \times 10^{-5}$ | 0.684057 |
| MPAATOT | $4.71 \times 10^{-8}$ | $5.4 \times 10^{-10}$ | $4.46 \times 10^{-5}$ | 0.12438 | 0.000129 |
| Critic review | $5.21 \times 10^{-9}$ | 0.006139 | 0.26457 | 0.05473 | 0.018521 |
| Budget | $< 2 \times 10^{-16}$ | $2.0 \times 10^{-8}$ | $< 2 \times 10^{-16}$ | $< 2 \times 10^{-16}$ | $1.2 \times 10^{-11}$ |
| Adjusted $R^2$ | 0.5291 | 0.3755 | 0.5820 | 0.5692 | 0.4711 |
| *N*(movies) | 671 | 165 | 206 | 141 | 159 |

Table 1 (above): MLR results with GENRETOT, MPAATOT, online critic review rating, and production budget predicting the total gross of movies are shown for each cluster. In each column the *p*-values for the correlation of each the four variables and the intercept are shown. The *p*-values represent the significance of the corresponding feature being non-zero. A significance level of $p < 0.05$ was used. Adjusted $R^2$, and the number of movies in the cluster are also shown for each cluster.

## Results (cont.)

GENRETOT, MPAATOT, online critic review rating, and production budget were used to predict the total gross of movies through MLR. GENREAVG, MPAAAVG, SEASON, and SLEEPERSCALE were all excluded due to dependencies and/or insignificant performance. When performed on the un-clustered data, all four variables were significant. Despite this, the overall model was still a poor predictor of a movie's total gross, as reflected in the low adjusted $R^2$ value (Table 1). Separating the data with *K*-means clustering (Graph 1) prior to performing MLR yielded no significant changes in adjusted $R^2$ values, only minimal decreases or increases, however it did result in some variables losing significance in certain clusters (Table 1). Overall, *K*-means clustering made no significant improvement upon the model's poor predictive value.

## Conclusion

This study was unable to find a robust correlation between in-theater movie competition conditions and box office sales through the use of *K*-means clustering and MLR. The predictive model was unable to account for the data's large amount of variance. The movie data proved to be excessively noisy. There lacked any clear distinctions in the data that the *K*-means clustering could meaningfully analyze, resulting in no improvement in the model when it was applied. Future studies should approach the *K*-means clustering with more than two dimensions, since two proved to be insufficient in finding clusters. There are many more factors at work in determining the box office success than could be analyzed in the time allotted for this study. One such factor being theater distribution/release strategy. Movies that are shown in more theaters naturally have more exposure and thus may have higher gross, but may also end up having a shorter lifespan. Analysis could also be focused towards opening week as it most often is the largest portion of a movie's gross. Competition in the opening week may be much more significant than competition halfway through the movie's lifespan. Competition generated outside of the theater could also be examined. Research could explore competition generation effects of movie advertisements.

## References

Jedidi, K., Krider, R., & Weinberg, C. (1998). Clustering at the movies. *Marketing Letters, 9*, 393-405. Retrieved from http://www.jstor.org/stable/40216182