The Study of the Beauty of Film

Survey Design and Implementation Team Project

Gurjus Singh, Ashton Duke and Maria Lopez Gonzalez

Northwestern University

Introduction

Studios have to conscientiously think about ways to stay competitive in an industry where there is ongoing development in movie-making. To gain a competitive advantage, they use past history such as budget, storyline, and year to analyze trends on what genres will garner high profits (Vincent, 2019). We are a survey group hired by a select number of studios that are studying popular movies and what the audiences of these movies think about them. We will determine if these top preferred movies in the survey we will implement correspond to the box-office earnings of the movies Our prediction is that individual preferences will strongly correlate with movie box-office performance.

Studies like this have been done within the movie industry and they used similar methods to predict movie performance in the box office. One study used brain waves from people watching movie trailers to movie success (Christoforou et al., 2017). This study showed that neuroscience methods such as eye-gazing and electroencephalogram can be used successfully in predicting a movie’s success (Christoforou et al., 2017). Another study used sentiment analysis in predicting movie ratings (Jain, 2013). The movie’s performance was analyzed using viewer’s tweets about the movie to predict box office earnings (Jain, 2013). Specifically, different moods were extracted from the tweets about the movies (Jain, 2013). A third study tried to connect how political affiliation affected preferences on movies (RSM Discovery, 2014). In this study, it was found that “Democrats tended to prefer movies with African-Americans males in lead roles while Republicans tended to prefer movies with lead roles filled by Caucasians females” (RSM Discovery, 2014).

Description of population and sample

For the population, our sampling frame will be people who are members of the theater clubs. Specifically, we will look for movie preferences in order to predict box office earnings since people that spend money at the movie theater greatly impact the earning reports. We will be able to collect our results by partnering with movie clubs in different movie theaters such as Cinemark Movie Club or Regal Unlimited to collect a list of their members who are already associated with watching the movies to take our survey (Cinemark, n.d.; Regal, n.d.).

The sampling method that will be the most useful in our case is a stratified random sample (Hayes, 2020). Stratified random sampling will be particularly useful because it involves organizing groups who have watched the relevant movies from our list in the theaters (Hayes, 2020). We will divide the population into 3 groups depending on who has seen multiple of the relevant movies in the theater, who has only seen one relevant movie, or who has seen none of the movies listed in our survey. We will then choose randomly from these groups of people, and see each group's impact on movie earnings.

Discussion of Survey

In our study, we are using people to predict box-office performance by using techniques such as linear regression, clustering and sentiment analysis. We will ask them for their political views, age, gender, and income level. If people are uncomfortable with our survey, they can decide to opt out of specific questions. People can only fill out the survey once. There will be no incentives for taking the survey, but the survey will be beneficial to movie watchers who want to see better movies, because we will be able to provide information to studios on what movies people would like to see. We will ask them about their movie preferences and see which movies they prefer the most out of the list of movies which we generated based on popularity.

The data for box-office earnings has already been collected. We plan to learn if these preferences have anything to do with their favorite movie category. We will also ask them to elaborate on their rankings through a free response question. Our hypothesis is that the top movie preferences will indeed lead to movies that have higher box-office earnings.

Discussion of sources of error & bias

One bias that might result from our sample and study is the socioeconomic class of people that actually go to the theater. This bias is because most moviegoers are from the middle and upper class, so low-income households might not be heavily weighted in this sample. Also, there might be errors and respondent bias when identifying gender, income level, and political views because some people in this study would not be comfortable releasing how much they make, what gender they identify as or what their political views they have. Some groups might be underrepresented more than others, so this might be another cause bias in the box-office performance as well.

Proposed methods of data analysis

For data cleaning purposes, we will check if the data is consistent and there are no missing values in the fields. We will remove the data that does not fit these standards before performing our analysis. For example, we might have to perform imputation techniques on data that do not include political views, gender, and/or income class. During data cleaning we will use descriptive statistics to find outliers in the data. For example, we can box and whisker plots to find these outliers. These outliers can be the representation of our missing data that we can do imputation on.

After we have done post-survey data cleaning, we will either use linear or logistic regression, depending on our how data looks, to see which movies tend to predict higher box office performance. We will also look to see from the people that took are survey, and saw the movies in the theater, whether their data corresponds to accurate predictions in box office performance since these types of people greatly impact movie performance. Regression algorithms in our case are important because we want to see if preferences in natural language can be predicted by these algorithms and how accurately the predictors can perform (Statistics Solutions, 2020).

We will also consider the use of other algorithms such as clustering, which can be used to predict common traits between the top movies in our survey. This can be a clustering type problem because we already have the data available from the studios about each movie in our survey including genre, budget, year. We can identify common attributes such as budget, category and year among each movie using clustering as it groups common movies together (Matteucci, 2020). We can also use clustering to see which groups from our sample as the most impact in movie earnings and performance. The clustering algorithm that we can use is k-means since it groups data into clusters the researcher wants and calculates the distance between each cluster and each data point (Matteucci, 2020).

Lastly, we can use sentiment analysis, to analyze each mood that the respondent uses to describe the movie, and their experience with it. Sentiment analysis is important to our research as it describes the mood of a piece of writing. In this case, we can create a list of common attitudes to sort answers on the free response questions in our survey and describe their views in a neutral, positive or negative way (MonkeyLearn, 2020). We can then find the average rating of these films based on average rankings calculated from each person by selecting a bin such as 1-10 top ranked movies as positive, 11-15 range as neutral, and 16-20 range as negative and put films in these correct bins. Finally, we will add the films' mood back to our movie data set for final cluster analysis as another attribute column.

Conclusion

With our research we are hoping to find how individuals can reflect movie box performance. We are focusing on how people's individual rankings can add up to a movie performing well in the box office. Our expected value is by averaging people's rankings and preferences that movies that have a higher preference will perform better at the box-office. This will reflect in the data that we collected on top performing movies.

References

Cinemark. (n.d.). Cinemark movie club: the movie-lover's membership. Retrieved March 4, 2020, from https://www.cinemark.com/movieclub

Christoforou, C., Papadopoulos, T. C., Constantinidou, F., & Theodorou, M. (2017, November 30). Your brain on the movies: a computational approach for predicting box-office performance from viewer's brain responses to movie trailers. Retrieved February 19, 2020, from https://www.frontiersin.org/articles/10.3389/fninf.2017.00072/full

Hayes, A. (2020, February 19). reading into stratified random sampling. Retrieved March 4, 2020, from https://www.investopedia.com/terms/stratified_random_sampling.asp

Jain, V. (2013, March). Prediction of movie success using sentiment analysis of tweets. Retrieved February 19, 2020, from http://www.jscse.com/papers/vol3.no3/vol3.no3.46.pdf

Matteucci, M. (n.d.). Clustering: an introduction. Retrieved February 27, 2020, from https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/

MonkeyLearn. (n.d.). Sentiment analysis. Retrieved February 27, 2020, from https://monkeylearn.com/sentiment-analysis/

Regal. (n.d.). Regal Unlimited movie subscription pass. Retrieved March 4, 2020, from https://www.regmovies.com/static/en/us/unlimited

RSM Discovery. (2014, May 2). Political preference predicts movie choice. Retrieved February 27, 2020, from https://discovery.rsm.nl/articles/detail/113-political-preference-predicts-movie-choice/

Statistics Solutions. (n.d.). What is linear regression? Retrieved February 27, 2020, from https://www.statisticssolutions.com/what-is-linear-regression/

Vincent, J. (2019, May 28). Hollywood is quietly using AI to help decide which movies to make. Retrieved March 8, 2020, from https://www.theverge.com/2019/5/28/18637135/hollywood-ai-film-decision-script-analysis-data-machine-learning

Appendix

Movie Preferences Survey

Instructions: Please fill out this survey to help us gain insight on your movie preferences.

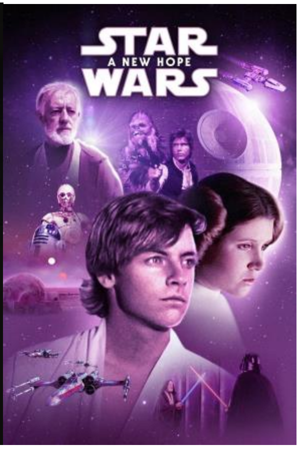
1. Select your age range.
    a. 18-24
    b. 25-34
    c. 35-44
    d. 45-54
    e. 55-64
    f. 65 or more
    g. Prefer not to answer
2. Select your gender identity.
    a. Male
    b. Female
    c. Other
    d. Prefer not to respond
3. How often do you go to movie theaters?
    a. Often
    b. Occasionally
    c. Rarely
    d. Never
4. Do you like buying the DVD/BLU-RAY release of the movie?
    a. Yes
    b. No
5. Do you have a streaming service at home?
    a. Yes
    b. No
6. If you answered yes to question 5, please list what the streaming services are.
7. Would you consider yourself:
    a. Lower Class
    b. Middle Class
    c. Upper Class
    d. Prefer not to say
8. Which political party best aligns with your political views?
    a. Democratic
    b. Republican
    c. Independent
    d. Other
    e. Prefer not to say

9. Rank these films from 1-20 in a preference you prefer. A list of all of the movies can be found after the series of movie posters.

1.

JURASSIC PARK

2.

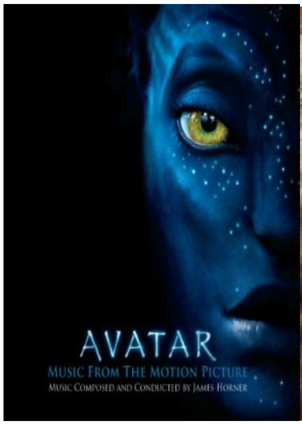


3.

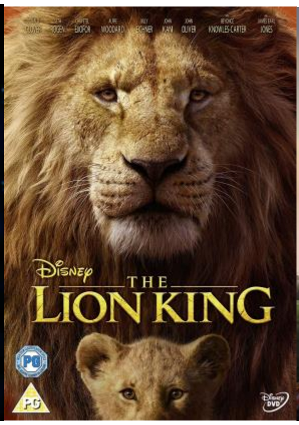
4.

NOTHING ON EARTH COULD COME BETWEEN THEM
LEONARDO DiCAPRIO KATE WINSLET
TITANIC

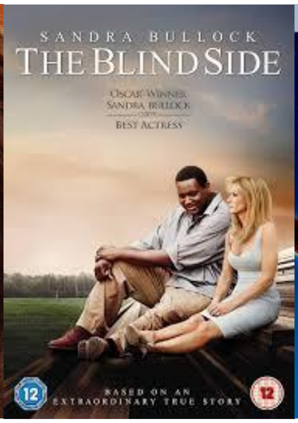
5.


AVATAR
MUSIC FROM THE MOTION PICTURE
MUSIC COMPOSED AND CONDUCTED BY JAMES HORNER

6.



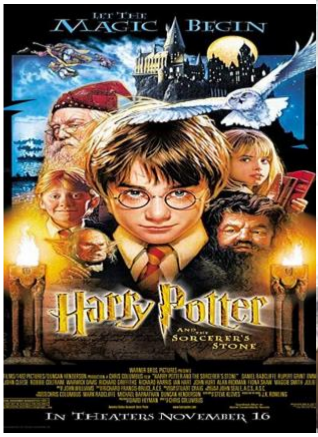
7.


SANDRA BULLOCK
THE BLIND SIDE
OSCAR WINNER SANDRA BULLOCK BEST ACTRESS
BASED ON AN EXTRAORDINARY TRUE STORY
12

8.

ANNIVERSARY EDITION
BLU-RAY + DVD + DIGITAL COPY + ULTRAVIOLET
A STEVEN SPIELBERG FILM
E.T.
THE EXTRA-TERRESTRIAL

9.

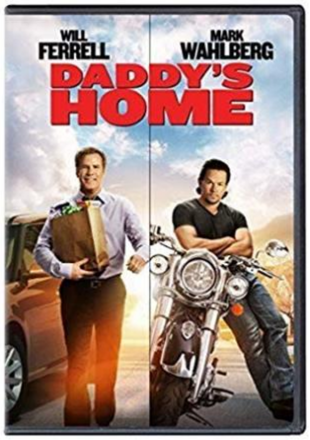


10.

11.

MARVEL STUDIOS
AVENGERS
ENDGAME
APRIL 26

12.



13.

LINDSAY LOHAN
MEAN GIRLS
WATCH YOUR BACK
IN THEATRES APRIL 30
AND TINA FEY

14.

TOM CRUISE
MISSION: IMPOSSIBLE
FALLOUT
SOME MISSIONS ARE NOT A CHOICE
JULY 27

15.

DANIEL DAY-LEWIS
LINCOLN
NOVEMBER

16.

THE ART OF THE FILM
WORLD WAR Z
BASED ON THE NOVEL BY MAX BROOKS

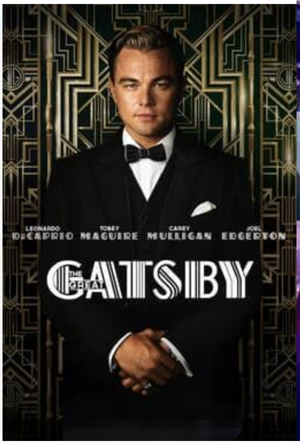
17.

FROM THE DIRECTOR OF SAW AND INSIDIOUS
THE CONJURING
BASED ON THE TRUE CASE FILES OF THE WARRENS
COMING SOON

18.

BLU-RAY + DVD + DIGITAL CODE
ANNABELLE
COMES HOME
"THE MOST HAUNTING ANNABELLE YET."

19.



20.

HUSTLERS
CONSTANCE WU
JENNIFER LOPEZ
JULIA STILES
KEKE PALMER
LILI REINHART
LIZZO
AND CARDI B
INSPIRED BY A TRUE STORY
IN THEATERS SEPTEMBER 13

List of Movies Above

1. Jurassic Park (1993)
2. Star Wars A New Hope (1977)
3. Hidden Figures (2016)
4. Titanic (1997)
5. Avatar(2009)
6. The Lion King (2019)
7. The Blind Side (2009)
8. E.T. (1982)
9. Harry Potter and the Sorcerer's stone (2001)
10. Ted (2012)
11. Avengers: Endgame (2019)
12. Daddy’s Home (2015)
13. Mean Girls (2004)
14. Mission Impossible Fallout (2018)
15. Lincoln (2012)
16. World War Z (2013)
17. The Conjuring (2013)
18. Anabelle Comes Home (2019)
19. The Great Gatsby (2012)
20. Hustlers (2019)

10. Please describe in words how you came up with this ranking. Describe your first pick in words such as adjectives and negative/positive words. Describe if anything influenced you such as family background

____

____

____

____

____

11. Please choose your favorite movie genre

f. Adventure
g. Comedy
h. Drama
i. Horror
j. Musical
k. Romantic Comedy
l. Thriller Suspense

Thank you so much for your feedback. We hope we can enhance your future movie experiences with the feedback you provided.