

Quantitative Study of the Movie Industry Based on IMDb Data

by

Kanika Almadi

Master of Science, Embedded Systems
Nanyang Technological University, Singapore

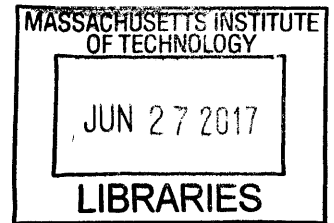
Submitted to the MIT System Design and Management Program
in partial fulfillment of the requirements for the Degree of

Master of Science in Engineering and Management
at the
Massachusetts Institute of Technology

June 2017

© 2017 Kanika Almadi. All rights reserved.

The author hereby grants to MIT permission to reproduce and to distribute publicly
paper and electronic copies of this thesis document in whole or in part in any medium
now known or hereafter created.



ARCHIVES

Signature of Author: _____ **Signature redacted** _____
Kanika Almadi
System Design and Management Program

Certified by: _____ **Signature redacted** _____
T. Tony Ke
Assistant Professor of Marketing, MIT Sloan School of Management

Certified by: _____ **Signature redacted** _____
Clair ZQ Yang
Postdoctoral Fellow, MIT Sloan School of Management

Accepted by: _____ **Signature redacted** _____
Joan S. Rubin
Executive Director, System Design and Management Program



77 Massachusetts Avenue
Cambridge, MA 02139
<http://libraries.mit.edu/ask>

DISCLAIMER NOTICE

Due to the condition of the original material, there are unavoidable flaws in this reproduction. We have made every effort possible to provide you with the best copy available.

Thank you.

The images contained in this document are of the best quality available.

THIS PAGE IS INTENTIONALLY LEFT BLANK

Quantitative Study of the Movie Industry Based on IMDb Data

by

Kanika Almadi

Submitted to the MIT System Design and Management Program in partial fulfillment of the requirements for the Degree of Master of Science in Engineering and Management

Abstract

Big Data Analytics is an emerging business capability that is providing far more intelligence to the companies nowadays to make well-informed decisions and better formulate their business strategies. This has been made possible due to easy accessibility of immense volume of data stored in clouds in a secure manner. As a result, online product review platforms have also gained enormous popularity and are successfully providing various services to the consumers primarily via user-generated content.

The thesis makes use of raw and unstructured data available on IMDB website, cleans it up and organizes it in a structured format suitable for quick analysis by various analytical softwares. The thesis then examines the available literature on analytics done on IMDB movie dataset and identifies that little work has been carried out in predicting the financial success of the movies. The thesis thus carries out data analytics on the IMDB movie sets and highlights several parameters like movie interconnectedness and director's credentials, which correlates positively with the movie gross revenue.

The thesis thereafter loosely defines a movie innovative index encompassing of parameters like number of references, number of follows and number of remake and discusses how the abundance of some of these parameters have a positive impact on box office success of the movie. Contrarily the lack of presence of these parameters thereby characterizing an innovative movie may not be so well received by the audiences thus leading to poor box office performance. The thesis also proposes how the director's credentials in the film industry measured by his/her total number of nominations and awards winning in the Oscar have a positive impact on the financial success of the movie and their own career advancement.

Thesis Advisors

T. Tony Ke
Assistant Professor of Marketing
MIT Sloan School of Management

Clair ZQ Yang
Postdoctoral Fellow
MIT Sloan School of Management

THIS PAGE IS INTENTIONALLY LEFT BLANK

Acknowledgements

I would like to thank my Thesis advisor Professor Tony Ke for providing me an opportunity to work under his guidance, in a research-motivated environment. I have learned a lot from him during both in class teaching and during thesis discussion and would always be grateful to him for his insightful ideas and valuable suggestions regarding the thesis execution.

I would like to express my sincere gratitude and thanks to Clair Yang. Clair's guidance at the later stages of my thesis was invaluable. She has been instrumental in defining the thesis scope, helping me out with the requirements and contributing to the research work and I thank her for kindly agreeing to be my Thesis Co-Advisor with such a short notice.

I would also like to thank Pat Hale, Joan Rubin, and William Foley from SDM department for guiding and supporting me patiently throughout my studies at MIT.

Finally, I would like to thank my family, for without their loving support, I would not be able to complete this educational journey.

THIS PAGE IS INTENTIONALLY LEFT BLANK

Tables of Contents

Tables of Contents.....	7
List of Figures.....	8
List of Tables	9
Chapter 1 Introduction.....	10
1.1 Motivation.....	10
1.2 Methods.....	11
1.2.1 Data Extraction, Cleaning, Transformation and Linking.....	11
1.2.2 Data Analytics	13
1.3 Result Summary.....	14
Chapter 2 Literature Review	16
2.1 Literature of Innovation.....	16
2.2 Literature of Motion Picture Industry.....	18
2.3 Literature of Online Platforms	21
2.3.1 IMDB.com.....	21
Chapter 3 Data	26
3.1 Data Sources.....	26
3.1.2 Wikipedia.com.....	34
3.1.3 Kaggle.com.....	35
3.2 Data Explanation	36
3.3 Summary Statistics	38
Chapter 4 Results.....	40
4.1 Financial Success and Interconnectedness of Movies.....	40
4.2 Effect of Past Academy Awards Nominations on Director’s Ability to make Future Movies .	41
Chapter 5 Conclusion	44
5.1 Significance.....	44
5.2 Future Areas of Work	45
References.....	47

List of Figures

Figure 1.1 Data Generation Process	12
Figure 2.1 Value Chain of Motion Picture [1]	19
Figure 2.2 Film Production in Different countries as of 2015 [2]	20
Figure 2.3 IMDB Homepage	22
Figure 2.4 Movie characteristics in the giant component	24
Figure 3.1 IMDB Movie Page	27
Figure 3.2 IMDB Movie Information Data source [A]	29
Figure 3.3 Example of Movie Scene Referenced in	37

List of Tables

Table 2.1 Summary Statistics of Econometric Analysis on Modern Art Creation [9].....	17
Table 2.2 Correlation among year –normalized financial data, user rating and user votes	25
Table 3.1 Movie.csv Data Format.....	32
Table 3.2 Reference.csv Data format	33
Table 3.3 Writers.csv Data format	33
Table 3.4 Directors.csv Data format.....	34
Table 3.5 Awards.csv Data format	36
Table 3.6 Summary Statistics.....	39
Table 4.1 Regression Analysis of Financial Success and Movie Interconnectedness.....	41
Table 4.2 Regression Analysis of Academic awards and various parameters.....	43

Chapter 1 Introduction

Internet has become ubiquitous now days. With the advent of electronic devices and affordable prices, we are always connected to the world and can access and share information easily on the online platforms. This lead to the popularity of numerous online movie platforms which consist of a database of all the movies that have been produced and released worldwide along with user ratings, reviews, awards information. IMDB is one of those online platforms, which is frequently accessed by the users all over the world and is one of the largest database containing information regarding movies, television programs and video games. Registered users on the IMDB website can contribute content voluntarily to the platform with no monetary benefit associated with it. Users are also provided with the option of rating the movie or a TV series on a scale of 1 to 10 with 10 being the highest. IMDB then calculate the overall movie rating based on the numbers of votes the movie has received and display it to the users.

1.1 Motivation

IMDB has always been a subject of scientific analysis because of the plethora of information available on its platform and easy access to large data sets for big data analysis. In the past, some studies have been done to figure out correlation between user voting and movie budget with the box office success of the movie [8]. However, little work has been carried out in the field of movies interconnectedness and financial success of movies as well as the effect of awards winning and academy award nomination on the director's ability to make more movies and generate future financial success. This motivated us to explore further in this field in order to get answers to these questions. We identified few parameters namely number of follows, references and remakes under connection type in IMDB dataset, which can potentially be grouped together to provide an estimate

of innovation index of the movie. Our assumption is that the higher the number of follows, references and remake for any movie, the less innovative that movie potentially is as most of the movie plot is matched with the other movie plots. We conducted data analysis on 966,431 movies present in IMDB dataset from year 1878 to 2016 and found out that there seems to be a positive correlation between gross revenue earned by a movie and its interconnectedness with other movies, namely number of follows, references and remakes. Thereby this suggests that box office success is defined by movies that are less innovative in nature. We also identified correlation between the academy award nomination for the best director in the previous year to the number of movies directed in the following year as well as the financial success associated with it.

1.2 Methods

In order to carry out the work on IMDB dataset, we first extracted, cleaned and arranged the data in the desired format and then conducted analytics on the transformed data using STATA software. The following two subsections explain in details about the methods used.

1.2.1 Data Extraction, Cleaning, Transformation and Linking

Although the IMDB list files are divided logically into separate .list files, the format of individual files is not suitable for analytics purposes. Hence, it is desirable to dump the files into a Relational Database (RDBMS) which allows to carry out big data analysis conveniently. We used MySQL Community Database Server for the said purpose as almost all the data analytics tools (R Studio etc) have MySQL plugins available. This allows the structured RDBMS data to be used as data source easily for the analytics program. We only parse specific .list files which have the data we need to conduct the analysis under scope of the project. While importing the data into MySQL DB, we also develop a relational information between various entities (actors, directors, writers, movies

etc). This relational information would allow us to easily import and analyze all the available information for a single entity while iterating through the list of various entities. For example, while looping through the list of movies, one could easily fetch all of its actors, directors, awards etc by writing simple SQL statements. In addition, data being available in MySQL Database allow us to quickly generate CSV flat files in varied desirable formats depending on the needs of the analytics program which is written in STATA (details of analysis are covered in later section).

To further illustrate the complete data transformation approach used, below flow diagram depicts the process pictorially:

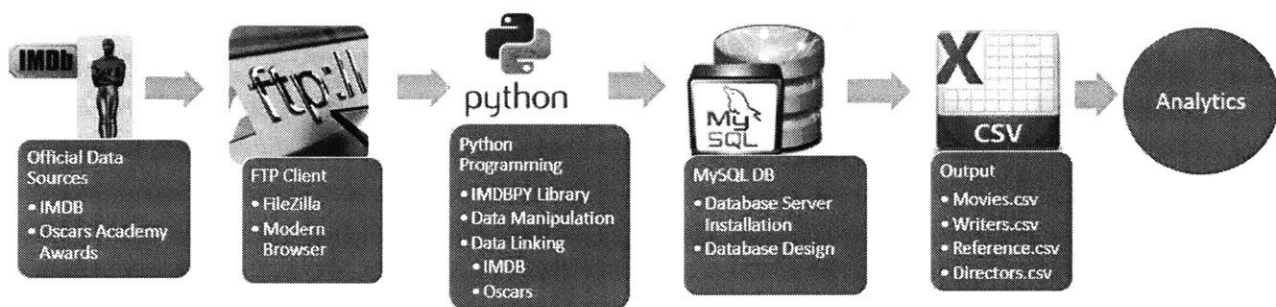


Figure 1.1 Data Generation Process

As shown in the above diagram, we make use of python as the programming language to generate the CSV flat files. Python is a high-performance programming language and is highly recommended in the field of data analytics. We deal with two different data sources in this project and make use of regex functions to link these data sources with 100% accuracy. Both the data sources are official data sources so the quality of data (and accuracy) is very good, although, we had to make some adjustments in the program to carefully handle some exceptional cases where there were no matches.

During the project execution, we derived multiple CSV flat files for various experimental purposes from the MySQL Database where as the processing of files from .list files (IMDB) and .csv files (Oscars) and loading into MySQL Database is one time activity. As mentioned earlier, linked and structured data being available in MySQL allowed us to generate various types of CSV files and carry out various experiments.

1.2.2 Data Analytics

After the data has been transformed into the desired format, we made use of the STATA software and performed correlation analysis and multiple linear regression analysis on the four csv files generated. Three types of linear regression analysis were carried out –

In the first regression equation - movie gross revenue was selected as a dependent variable and number of follows, references and remakes of a movie were listed as independent variables.

$$\mathbf{Gross\ Revenue}_i = \beta_1 \mathbf{Num\ Ref}_i + \beta_2 \mathbf{Num\ Follow}_i + \beta_3 \mathbf{Num\ Remake}_i + \alpha_t + \gamma_j$$

1

In the second regression equation – number of movies directed by a director in a particular year was selected as a dependent variable and number of movies that got nominated as well as number of movies that were awarded Oscars in the previous year were listed as independent variables.

$$\mathbf{Num\ Movie}_{kt} = \beta_1 \mathbf{Nomination}_{k,t-1} + \beta_2 \mathbf{Win}_{k,t-1} + \beta_3 \mathbf{X}_{kt} + \alpha_t + \gamma_j$$

2

In the third regression equation – total gross earned by the movie for a particular director was selected as a dependent variable and number of movies that got nominated as well as number of movies that were awarded Oscars in the previous year were listed as independent variables.

$$Total\ Gross_{kt} = \beta_1\ Nomination_{k,t-1} + \beta_2\ Win_{k,t-1} + \beta_3\ X_{kt} + \alpha_t + \gamma_j$$

3

1.3 Result Summary

- I. In the first regression analysis data samples were taken from three different categories -
 1. 2203 observations were taken for movies produced by US based company/production house with year as fixed effect.
 2. 2,379 observations were taken for movies having English as one of its official language with year as fixed effect.
 3. 2,546 observations were taken for all movies with year and country as fixed effect.

Following observations were obtained –

1. The coefficient of number of references for a movie is statistically significant on the gross revenue.
 2. The coefficient of number of follows for a movie is statistically significant on the gross revenue.
 3. The coefficient of number of remake for a movie is negative and statistically insignificant on the gross revenue.
- II. In the second regression analysis, data samples were taken from the following categories –
 1. 475,133 observations were taken for all directors with year and individual as fixed effect.

2. 3,606 observations were taken for directors having Academy award nomination with year and individual as fixed effect.

Following observations were obtained –

1. The coefficient of nomination for best director in academy award in the lag year is statistically significant on the total number of movies produced in the following year.
2. The coefficient of academy award winning of best director in the lag year is negative and statistically insignificant on the total number of movies produced in the following year.
3. The coefficient of director's years of experience is negative but statistically significant on the total number of movies produced.

III. In the third regression analysis, data samples were taken from the following categories –

1. 3,340 observations were taken for all directors with year and individual as fixed effect.
2. 471 observations were taken for directors having academy award nomination with year and individual as fixed effect.

Following observations were obtained –

1. The coefficient of nomination for best director in academy award in the lag year is statistically significant on the gross revenue in the following year.
2. The coefficient of academy award winning of best director in the lag year is statistically significant on the gross revenue in the following year.
3. The coefficient of director's years of experience is partially significant on the gross revenue under all directors' categories and negative and statistically insignificant under directors having academy award nomination category.

Chapter 2 Literature Review

In order to conduct a meaningful research on IMDB dataset, we review the literature extensively over spanning the three broad categories namely the motion picture industry, innovation and online platform. Following sub sections describe in details about these categories.

2.1 Literature of Innovation

One of the main aspect to which a well-reputed movie critique pays attention to when reviewing the movie is the innovativeness of the overall movie. This may consist of the use of futuristic technology to display certain action/day to day tasks, innovative story line that has never been thought of before, sophisticated display and video making technology that enhances the user experience and so on. Movie making style needs to undergo a series of constant innovation to keep the audiences hooked to TV screen all the time as well as to maximize the overall profit of the motion picture industry. Several studies have been carried out in the past, which explores the impact of innovation in a particular field to the degree of success achieved by the associated body.

Galsenson and Weinberg [9] did a similar study on the invention of Modern art due to the constant demand for innovation and the corresponding success it has brought in the life of painters. They pointed out that the increased demand for innovation prompted the beginning of modern painting in which the artists made the painting a more of conceptual activity where age plays no role and young artist could as well make significant advances. They examined the relationship between artist ages and the market value of their paintings in the twentieth –century.

Galsenson and Weinberg conducted a data and econometric analysis on all painters born between 1830 -1900 and who were natives of France and had created at least one painting which had been

reproduced in at least three out of the five designated textbooks of art history. They divided the total artists by their born year into 4 different cohorts (1820 – 1839, 1840 – 1859, 1860 – 1879, 1880 – 1900). They took the sample of total 13,943 paintings out of which 11 percent of the paintings were created by artists in the first cohort, 15 percent of the paintings by second cohort, 31 percent by third cohort and 43 percent by fourth cohort.

The following table shows the summary statistics of the data.

	All	1820–1839	1840–1859	1860–1879	1880–1900
Year of birth	1869 (20.0)	1833 (3.4)	1842 (4.0)	1873 (4.7)	1886 (5.8)
Year of execution	1920 (26.3)	1885 (11.8)	1893 (13.3)	1919 (16.2)	1940 (19.2)
Age at execution	50.8 (16.7)	51.7 (12.3)	51.0 (14.6)	46.1 (14.9)	53.9 (18.7)
Year of sale	1985 (7.6)	1985 (7.9)	1985 (7.5)	1985 (7.5)	1986 (7.6)
Price	370,151 (1,459,712)	547,700 (1,313,422)	833,029 (2,654,701)	168,484 (528,846)	306,101 (1,297,916)
Area	527.2 (651.3)	407.6 (591.2)	463.4 (480.0)	522.3 (660.6)	583.2 (704.1)
Paper	0.315 (0.465)	0.430 (0.495)	0.108 (0.311)	0.289 (0.453)	0.381 (0.486)
Observations	13,943	1468	2156	4343	5976
Number of artists	33	4	7	8	14
Paintings per artist	423	367	308	543	427

Table 2.1 Summary Statistics of Econometric Analysis on Modern Art Creation [9]

They conducted a linear regression for examining the relationship between price of the painting and the age of the artist at the date of the painting execution. Regression was carried out in which the natural log of real sale price of the painting was expressed as a polynomial consisting of artist age as interacted with its birth cohort.

Their analysis reflected that due to an increased demand of innovation and influence of modern art over time, more young painters contributed towards modern art by incorporating the idea of conceptual painting and discarding the traditional methods. Thus, number of years of experience

of painters hold no significance towards greatest artistic achievements as the young artists made most of the dramatic advances in the modern art industry.

The above findings intrigued our interest for conducting a similar study in the motion picture industry and exploring the relationship among director age/years of experience in movie directing, his/her contribution to movie innovation and the box office success associated with it.

2.2 Literature of Motion Picture Industry

Motion picture industry have always been of high economic importance towards the global economy because of its ability to provide employment to millions of people worldwide. In US alone, the industry employed over 406,000 people as stated by the Bureau of Labor Statistics 2016. In 2015, around \$11.1 billion of revenue was generated through the sale of theatrical tickets alone in North America and around \$27.2 billion of revenue was generated outside the North America. Motion picture industry also has a high cultural significance on different world's culture and weekly update of box office statistics is a much-awaited activity by people around the world.

The industry primarily comprises of three players – producers, distributors and exhibitors. Below figure represents the value chain for Motion Pictures.

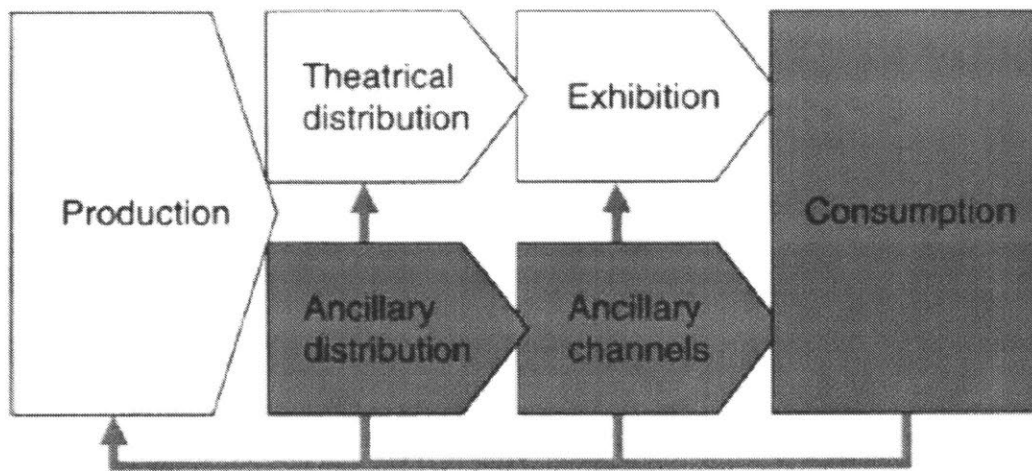


Figure 2.1 Value Chain of Motion Picture [1]

Producers /production companies are in charge of all development process related with movie making including story plot selection, recruitment of director, cast and crew, financing the movie budget and involvement in post-production activities like editing, dubbing, music selection/addition and finally before official release getting the approval of MPAA(Motion Picture Association of America) for movies produced in United States. India and United states are the largest producers of feature films and as of 2011, India produced around 1255 movies and United States produced 819 as cited by Wikipedia. As of 2016, global box office collection of all films released worldwide reached \$38.6 billion [3], with US/Canada box office collection reported to be \$11.4 billion.

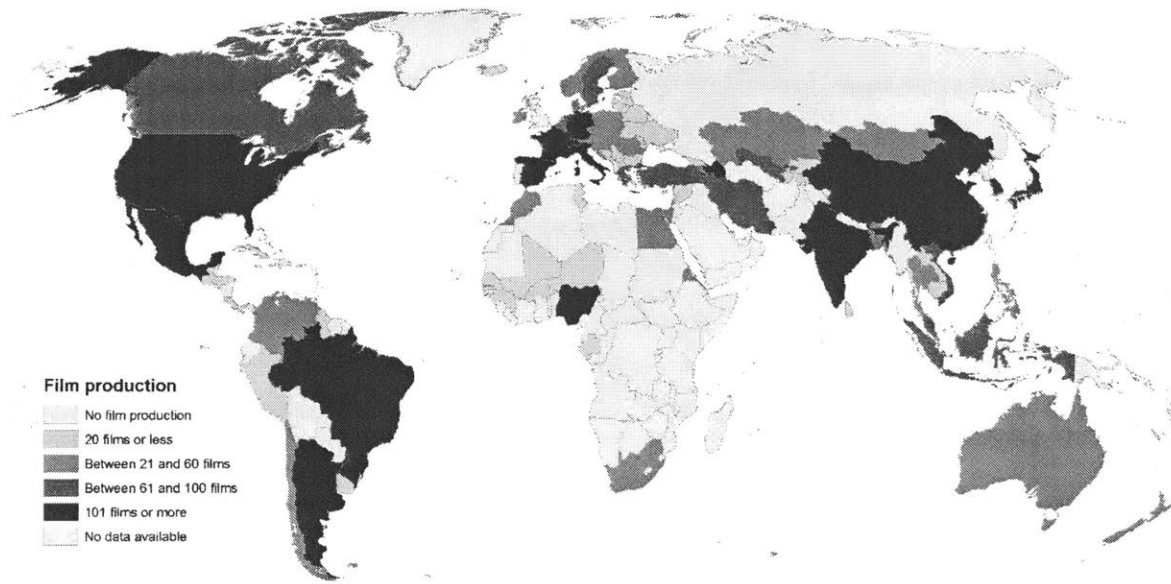


Figure 2.2 Film Production in Different countries as of 2015 [2]

Distributors are responsible for nationwide distribution of the completed movie and make decisions regarding selection of release date, release locations, contract negotiations with exhibitors as well as designing the nationwide advertisement campaign for movies. The strong seasonal effects in demand and the competition given by other movies throughout the movie run duration governs the release date selection of movies. Einav [4] studied the impact of seasonality on the motion picture industry of US and concluded that gross seasonality is amplified by the movie release decisions whereas underlying demands only account for two-thirds variation of the total sales. His findings also reflected endogeneity of observed seasonal patterns.

Exhibitors are primarily the theatre owners, who display the movie to audiences, and are not vertically integrated with either distributors or producers.

As studied above, Motion picture industry is one of the primary source of entertainment worldwide and has a strong ability to create a financial impact on the lives of millions of people, thus it

intrigued our interest to conduct some further research in this field to be able to add value to its economic and social contributions.

2.3 Literature of Online Platforms

Consumers have always been excellent marketers because they have used the product and can offer more personalized and relevant review about the product. As a result, review websites started gaining enormous popularity where reviews about people, products, businesses or services can be easily posted at the comfort of consumer's time. These review websites earn their revenue via advertisements where different businesses can list their products without affecting the actual reviews and ratings. These online platforms provide the option to post reviews anonymously as well, which ensured the unbiased feedback regarding any product. Shoppers on the other hand can easily fetch information shared by countless other consumers about products ranging from businesses, restaurants, movie reviews, physicians, etc.

Several studies have been carried out in the past using user generated content which aims to study the impact of customers review on the product demand. Luca [5] conducted a similar study using Yelp customer review data about restaurants and identified a causal impact of the restaurant rating on Yelp platform on the demand. He concluded that an increase in one star rating on the platform leads to an overall 5-9 percent increase in the restaurant revenue. This effect is observed in independent restaurants, and the ratings do not have effect on restaurants with chain affiliation. He also observed that there is decline in market share of chain restaurants with the increase in Yelp penetration.

2.3.1 IMDB.com

The Internet Movie Database (IMDB) was launched in 1990 and is one of the largest online database containing information about movies, television programs, video games, biographies of cast and crew of movies, movie reviews and plot summary. As of April 2017, IMDB has 75 million registered users along with 4.2 million titles and 7.8 million personalities records in its database [6]. IMDB as a subsidiary of Amazon Inc generates its revenue through advertisements, partnerships and licensing. Registered users can voluntarily contribute content to the platform. However, the submitted data undergo a series of consistency checks before being displayed on the platform and the contributors cannot add, delete or modify the data on impulse, which makes it different from other user-contributed websites like Wikipedia.com.

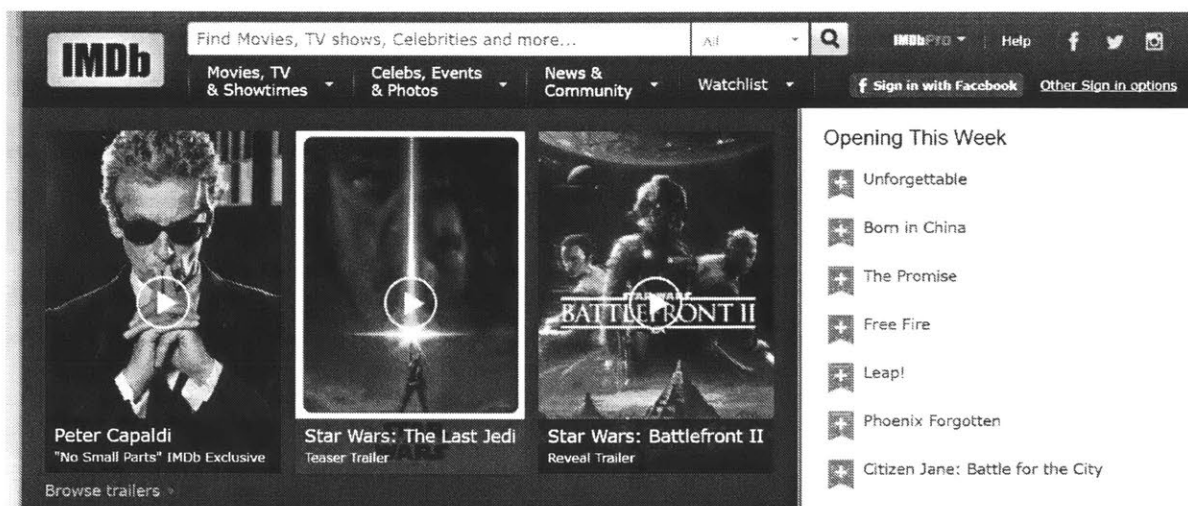


Figure 2.3 IMDb Homepage

IMDB dataset brings with it an element of interest for further study because most of the people are aware of the movie industry and the associated actors and when people are presented with a visualization of the movie dataset, most of them would try to look out for their favorite movies and actors and try exploring complex relationships among actors. IMDB dataset is stored in a very

clean and structured format and contains very rich information about each movie and its casting thereby allowing for a broad range of data analysis.

Wasserman and his team [8] conducted a study on IMDB data for films released between 1920 to 2011 and classify the films into three broad groups namely “USA” (Group 1) -Films Produced in USA regardless of language, “English non – USA” (Group 2)- Films made outside USA, with English as their primary language, “Non- English Non – USA” (Group 3) – Films made outside USA in language other than English. They constructed a network of 28,743 films and 74,164 arcs and presented the following findings.

As seen in the below figure, there has been significant increase in the number of new films released over a period of 1960 to 2010 in all the three groups as stated above. Also, an overall increase is observed in average out-degree parameter over a period of time till year 1999, because connections between any number of films can only flow backwards in time. Out- degree parameter represents an outgoing arc or connection between films.

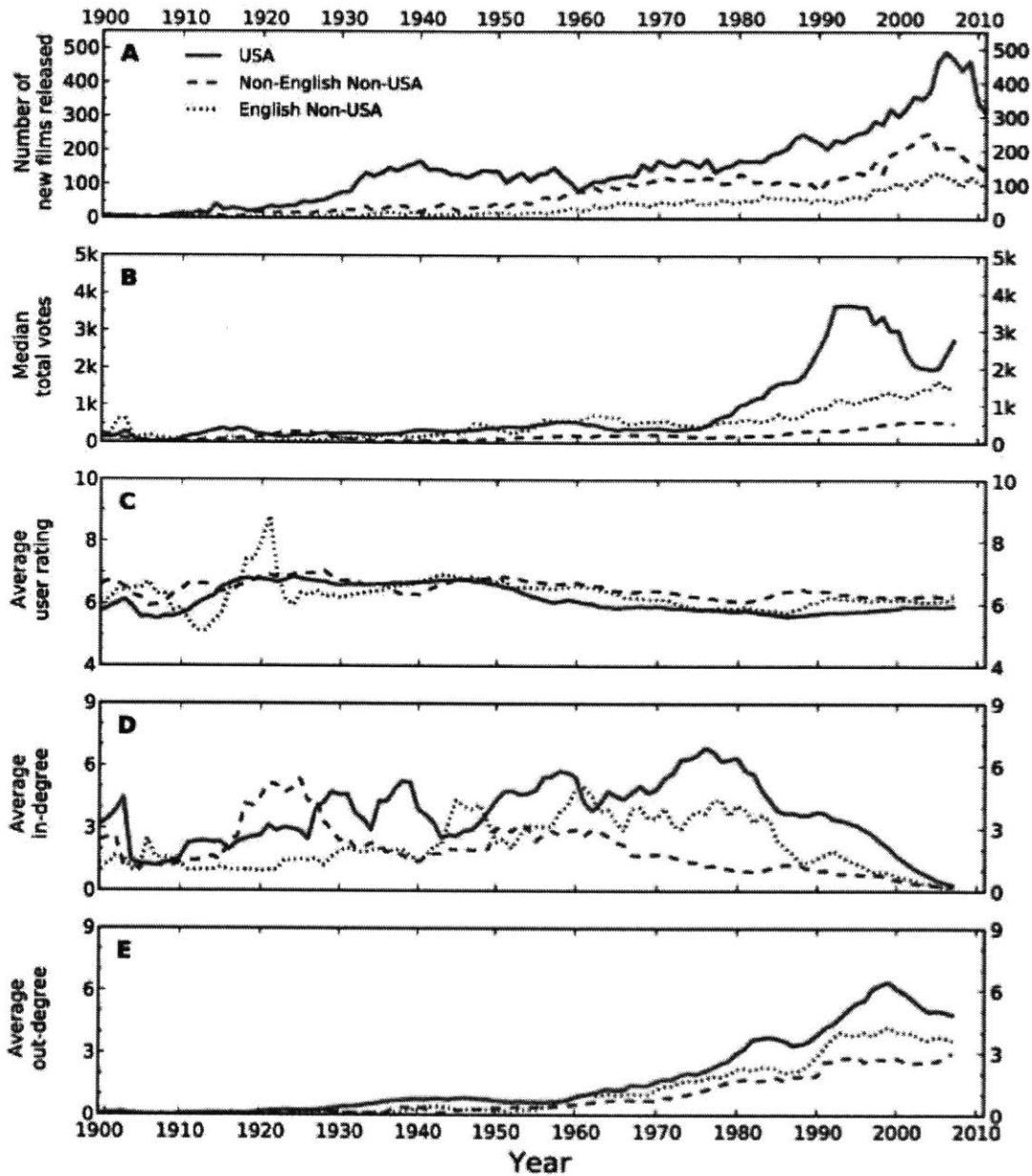


Figure 2.4 Movie characteristics in the giant component

Wasserman and his team presented us with an interesting correlation analysis between financial data namely the gross and the budget, user rating and user votes and observed a strong correlation between the log of normalized films budget and log of number of user votes. They concluded that

strong correlation among the two is due to high prominence of the film as more money is spent on the promotion of the film due to which more people become aware about the film and rate the film after viewing it. However, no correlation is observed between user rating and financial quantity.

The following table cites the result of correlation from Wasserman paper.

Model	r only		\hat{g}_r only		\hat{b} only		\hat{b} and r		All	
	Coeff.	SE.	Coeff.	SE.	Coeff.	SE.	Coeff.	SE.	Coeff.	SE.
Log of user votes										
Intercept	3.16	0.05	4.67	0.02	4.85	0.03	4.19	0.04	3.80	0.04
Log norm. budget \hat{b}	—	—	—	—	0.602	0.006	0.586	0.006	0.268	0.01
Log norm. total gross \hat{g}_r	—	—	0.354	0.006	—	—	—	—	0.191	0.008
User rating r	0.080	0.005	—	—	—	—	0.102	0.006	0.156	0.006
Probit selection										
Intercept	0.348	1.5	-80	2	-46	1	-46	1	-76	2
Year y	-5×10^{-5}	8×10^{-4}	0.040	8×10^{-4}	0.023	6×10^{-4}	0.023	6×10^{-4}	0.037	9×10^{-4}
Cube root in-degree $i^{1/3}$	1.19	0.07	0.916	0.02	0.713	0.02	0.713	0.02	0.884	0.02
Cube root out-degree $o^{1/3}$	0.915	0.06	0.354	0.02	0.250	0.02	0.250	0.02	0.317	0.02
Inverse mills ratio	-6.47	0.2	-0.81	0.02	-1.16	0.02	-1.12	0.02	-0.61	0.01
Total no. of films	15,425		15,425		15,425		15,425		15,425	
Films with observed data	14,577		5,307		5,331		5,331		3,430	
Films with censored data	848		10,118		10,094		10,094		11,995	
Adjusted R^2 %	26.02		63.74		73.97		75.50		72.83	

Table 2.2 Correlation among year –normalized financial data, user rating and user votes

The study done by Wasserman and his team motivated us to further explore the financial aspect of the movie industry and establish some positive association between movie gross revenue earned over its lifetime period and other factors included in the IMDB dataset which can affect its performance.

Chapter 3 Data

Availability of reliable data source was of prime importance for conducting our research and hence we made use of some of the popular databases available online whose data is either uploaded officially by the concerned authorities or verified by the trusted reviewers.

3.1 Data Sources

Following subsections describe about the various data sources used in this project.

3.1.1 IMDB.com

Internet Movie Database (IMDB) is one of the largest online databases for searching of information related to movies (long and short), TV series, video games etc. It also contains information about the casting characters of each movie, biography of people associated with entertainment industry, movie reviews, movie storyline etc. Most of the content is user generated and users are encouraged to voice their opinion by rating the movies, providing reviews and other useful information pertaining to the movies.

For each movie, the following information is present on the IMDB movie page –

- I. Movie Rating out of 10
- II. Movie Duration (hours & minutes)
- III. Genre (for example – Comedy, Drama, romance etc)
- IV. Movie Release date in each country
- V. Movie Director/s
- VI. Movie Writers
- VII. Movie Stars

- VIII. Storyline
- IX. Motion Picture Rating (MPAA)
- X. Movie Languages
- XI. Movie Filming Location
- XII. Box office information – Budget, Opening weekend, Gross, Weekend Gross
- XIII. Trivia
- XIV. Goofs
- XV. Famous Quotes
- XVI. Connections
- XVII. User Reviews
- XVIII. Academy awards, BAFTA, Golden Globes and other awards

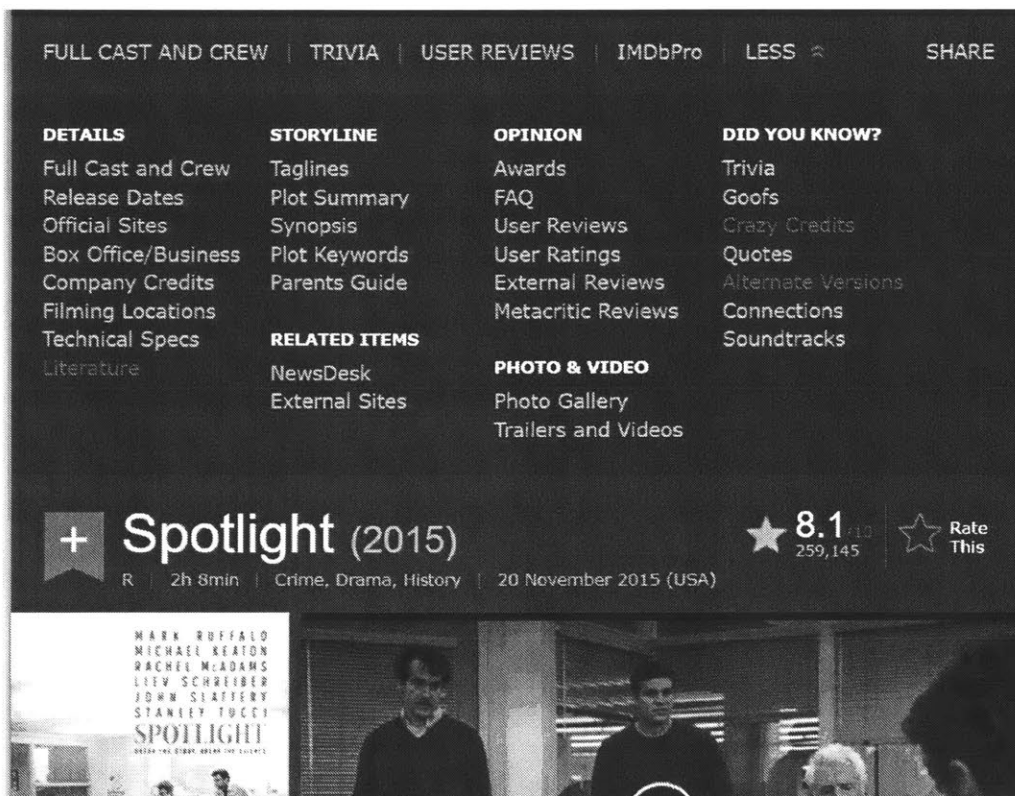


Figure 3.1 IMDB Movie Page

IMDB has provided two alternative interfaces besides the online version to access IMDb data locally by holding copies of the data directly on one's system. Data is for individual use under the IMDB license.

1. **The Plain Text Data Files** - The files are available for download from FTP sites. Any modern FTP client (or even a modern web-browser) is sufficient to download all the files. Most files carry **.list** extension. The data in these files uses the ISO-8859-1 character set, which is also known as Latin-1, with the exception of title data contained in the aka-titles.list. That data contains other character encodings, which are specified as attributes for the non-Latin-1 encoded entries.
2. **The Unix Command Line Search Programs** - Col Needham's package of Unix command line programs can be installed in the local machine. However, they are no longer supported by IMDb and may require manual modification in order to run.

We have use the plain text files for this project as it allows transforming the data in desired format.

Below figure is screen shot of IMDB data source.

Index of /pub/misc/movies/database/






























Name	Size	Date Modified
 [parent directory]		
 README	1.4 kB	5/29/14, 12:00:00 AM
 actors.list.gz	316 MB	3/24/17, 7:48:00 PM
 actresses.list.gz	178 MB	3/24/17, 7:52:00 PM
 aka-names.list.gz	8.5 MB	3/24/17, 8:06:00 PM
 aka-titles.list.gz	9.7 MB	3/24/17, 8:04:00 PM
 alternate-versions.list.gz	2.5 MB	3/24/17, 8:10:00 PM
 biographies.list.gz	205 MB	3/24/17, 8:04:00 PM
 business.list.gz	11.2 MB	3/24/17, 8:11:00 PM
 certificates.list.gz	6.0 MB	3/24/17, 8:06:00 PM
 cinematographers.list.gz	20.4 MB	3/24/17, 7:55:00 PM
 color-info.list.gz	18.5 MB	3/24/17, 8:07:00 PM
 complete-cast.list.gz	988 kB	12/19/15, 12:00:00 AM
 complete-crew.list.gz	580 kB	12/19/15, 12:00:00 AM
 composers.list.gz	16.1 MB	3/24/17, 7:55:00 PM
 contrib/		7/6/05, 12:00:00 AM
 costume-designers.list.gz	5.4 MB	3/24/17, 7:56:00 PM
 countries.list.gz	18.8 MB	3/24/17, 8:08:00 PM
 crazy-credits.list.gz	1.4 MB	3/24/17, 8:01:00 PM
 diffs/		3/25/17, 7:04:00 AM
 directors.list.gz	36.7 MB	3/24/17, 7:54:00 PM
 distributors.list.gz	28.3 MB	3/24/17, 8:11:00 PM
 editors.list.gz	26.2 MB	3/24/17, 7:56:00 PM
 filesizes	1.2 kB	3/24/17, 7:47:00 PM
 filesizes.old	1.2 kB	3/24/17, 7:47:00 PM
 genres.list.gz	18.2 MB	3/24/17, 8:06:00 PM
 german-aka-titles.list.gz	347 kB	12/19/15, 12:00:00 AM
 goofs.list.gz	23.1 MB	3/24/17, 8:02:00 PM
 iso-aka-titles.list.gz	20.8 kB	10/16/98, 12:00:00 AM

Figure 3.2 IMDB Movie Information Data source [A]

For the current project, we have made use of the following list files as the data source input from (<http://www.imdb.com/interfaces>), made use of the Python program to import the movie data into the SQL database, and generated the csv files containing the desired movie info for further analysis using a statistical tool.

[A] Information courtesy of IMDb (<http://www.imdb.com>). Used with permission.

IMDB List files used –

- I. Biographies.list
- II. Business.list
- III. Directors.list
- IV. Editors.list
- V. Genres.list
- VI. Movies.list
- VII. Movie-links.list
- VIII. Goofs.list

Following csv files were generated from the above list files after data transformation as described in section 1.2.1.

I. **Movie.csv**

Movie.csv contains total 966,431 rows and 28 columns. Each row in a csv contains information about movies represented by a unique movie id and data about biography, financial information, and connection of movies.

Below table lists all the variables present in the movie.csv file.

Number	Type	Explanation
1	Idm	Movie Unique id
2	Movie	Movie Name
3	Mpaa	Motion Picture Association of America movie rating
4	Runtime	Total movie runtime in hours & minutes

5	Country	Country of Production
6	Lan	Language of movie
7	Genres	Type of movie – eg comedy, romance
8	Release_date	Release date of movie
9	directors	ID's of the director/s who directed this movie
10	writers	ID's of the writer/s who wrote the movie script
11	Imdb_rating	IMDB user rating of the movie out of 10
12	Currency_budget	Currency type of the movie budget Eg Euro, USD, SGD
13	budget	Budget amount for the movie production
14	Currency_openweek	Currency type of the opening week revenue collected
15	Opening week	Total amount of revenue collected in the opening week
16	Currency_gross	Currency type of the gross revenue collected
17	Gross amount	Total amount of gross amount collected by the movie
18	Filming date –start	Filming date starting date and year
19	Filming date - end	Filming date ending date and year
20	Production companies	Production companies
21	Number of edits	Footage of tittle edited into subsequent movie

22	Number of edited	Some part of the movie is constructed by editing from previous movie
23	Number of references	Total number of references made by the movie
24	Number of referenced in	Total number of movies which referred the current movie
25	Number of features	Some extracts from previous movies are featured in current movie
26	Number of featured in	Some extracts from the current movie are featured in the subsequent movies
27	Number of spoofs	Total number of jokes references made to a previous movie
28	Number of spoofed	Total number of jokes of a current movie made in subsequent movies

Table 3.1 Movie.csv Data Format

II. Reference.csv

Reference.csv contains total 909890 entries where each row represents a connection between two different types of movies. The following table lists the format of the reference.csv file.

IDM1	IDM2	Type of Connection
Unique movie id 1	Unique movie id 2	<ul style="list-style-type: none"> • Follows • Spin off • References • Features • Referenced in • Spoofed in

		<ul style="list-style-type: none"> • Remake of • Followed by • Featured in • Edited from • Edited into • Alternate language version of • Version of • Spoofs • Remade as
--	--	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Table 3.2 Reference.csv Data format

III. Writers.csv

Writers.csv contains total 444,394 entries where each entry uniquely represents the biographical information about the movie writer and whether he worked as director as well or not. Below table, lists the format of the writers.csv.

Number	Type of column	Explanation
1	IDW	Writer's ID
2	Writer_Name	Writer name
3	Writer_DOB	Writer date of birth
4	Writer_place_of_birth	Writer place of birth
5	No_movies_written	Number of movies written by writer
6	Is_director	Is writer also worked as director

Table 3.3 Writers.csv Data format

IV. Directors.csv

Directors.csv contains total 365,113 entries where each entry uniquely represents the biographical information about the movie directors. Following table lists the format of the directors.csv file.

Number	Type of column	Explanation
1	IDD	Director ID
2	Director_Name	Director name
3	Director_DOB	Director date of birth
4	Director_place_of_birth	Director place of birth
5	No_movies_directed	Number of movies directed by director

Table 3.4 Directors.csv Data format

3.1.2 Wikipedia.com

Following Wikipedia links have been referred to for fetching information about the Academic awards won by the movies and directors from the year 1927 till 2015.

I. Academy award for Best Picture

The Academy of Motion Picture Arts and Sciences (AMPAS) presents this award annually to the producers of the films. It is considered as a highly prestigious and premium award because it takes into account all the effort put forth namely (acting, music composing, directing, writing, editing) into the movie production.

Web Source - https://en.wikipedia.org/wiki/Academy_Award_for_Best_Picture

II. Academy award for Best Director

The Academy of Motion Picture Arts and Sciences (AMPAS) presents this award annually to the directors of the movies for making an outstanding contribution towards film directing.

Web Source - https://en.wikipedia.org/wiki/Academy_Award_for_Best_Director

III. List of Academy Award – winning films

The following link lists all the movie names that have been nominated for the Academy awards as well as the total number of nominations and Academy awards won by each of those movies.

Web Source - https://en.wikipedia.org/wiki/List_of_Academy_Award-winning_films

3.1.3 Kaggle.com

In order to conduct the complete analysis, the academy awards data is required but sadly, this data is not provided by IMDB in flat file format. Hence, it is required to find alternative reliable data source to fetch the academy award data. The awards data has been made available to be downloaded for free by Academy of Motion Picture Arts and Science on popular data scientists forum Kaggle. Hence Kaggle data source was used along with Wikipedia source for obtaining information about the Academy awards information for the movies. Data was downloaded from the link (<https://www.kaggle.com/theacademy/academy-awards>) cleaned and reorganized using a Python program to generate awards.csv.

Below table is template of awards.csv generated for statistical analysis.

movie_id	movie	year	awards	nominations	nominated_best_picture	best_picture	nominated_best_director	best_director
3805863	The Noose	1928	0	1	NO	NO	NO	NO
3786144	The Last Command	1928	1	2	NO	NO	NO	NO
2752418	A Ship Comes In	1928	0	1	NO	NO	NO	NO
2733064	7th Heaven	1927	3	5	YES	NO	YES	YES
3603301	Sadie Thompson	1928	0	2	NO	NO	NO	NO
3756865	The Dove	1927	1	1	NO	NO	NO	NO
3719015	Tempest	1928	1	1	NO	NO	NO	NO

Table 3.5 Awards.csv Data format

As can be seen in the above table, each movie, which won the Oscar award, has been associated with a unique movie id, year of production, number of awards won by the movie in Oscars, number of nominations and a binary indicator of whether the movie won the best picture/best director award.

3.2 Data Explanation

Below is a list of some useful variables, which are used for data analysis.

i. References

It means that the current movie refers to some other movie scene/title or dialogue in a non-spoofed way.

For e.g. in movie Inception, the idea of using a machine to enter some other person mind in a dreaming state for solving a crime case has been referred from the movie Cell (2000).

ii. Referenced in

It means that the current movie title/dialogue or a scene is used or referred in a subsequent movie. For example, in movie -3 Idiots (2009)- the main characters give a titanic pose during a song sequence. This scene is imitated from the Titanic movie (1997).



Figure 3.3 Example of Movie Scene Referenced in

iii. Remade as

It means that a subsequent movie was made based on the current movie.

iv. Remake of

It means that the current movie is a remake of some existing movie.

v. Follows

It means that the current movie title is a sequel or follows a previous title released on an earlier date.

vi. Followed by

It means that the current movie title is followed by a subsequent movie/sequel.

vii. Gross (Worldwide)

For the gross variable, the financial gross data present in IMDB with the latest date and worldwide string attached to it is used.

viii. Opening weekend (Latest)

For the opening weekend, the financial opening weekend data present in IMDB with the latest date is used.

3.3 Summary Statistics

The following table provides the summary statistics of the data sample, including 966,431 movies and short films from the year 1878 to 2023 (as scheduled). It contains data on four categories: basic biographical information, financial information (such as opening week revenue and global gross revenue), Academy Awards information, and our connection index.

Summary Statistics						
Variable	Explanation	Obs	Mean	Std. Dev.	Min	Max
Number of movies		966,431				
Number of movies with USA as a production country		390,591				
Number of movies with English as a main language		483,262				
Number of feature movies		438,055				
Number of short films		528,376				
Number of movies with Academy Award Info		3,464				
Basic info						
release_year	release year	628,183	1984.119	36.52616	1878	2023
runtime	run time	615,876	46.16471	52.28538	1	14400
imdb_rating	IMDb rating	281,251	6.455551	1.379978	1	10
connection	number of inter-movie connections	966,431	0.633546	11.06974	0	4512
filming_length	length of filming dates (month)	26,998	2.694903	1.87544	1	12
Financial Info						
budget_usd2010	budget in constant 2010 USD	127,555	1.82E+11	4.87E+13	0	1.58E+16
budget_amt	budget in current LCU (local currency unit)	174,507	1.23E+09	5.03E+11	0	2.10E+14
gross_usd2010	gross revenue in constant 2010 USD	3,122	1.63E+08	2.17E+08	0	2.84E+09

gross_amt	gross revenue in current LCU	3,471	1.29E+08	1.84E+08	0	2.79E+09
openweek_screen	number of screens for opening week	9,049	441.7431	24465.65	1	2323323
openweek_usd2010	opening week revenue in constant 2010 USD	11,621	1.16E+13	1.25E+15	0	1.35E+17
openweek_amt	opening week revenue in current LCU	12,592	6.99E+12	7.84E+14	0	8.80E+16
Award info						
win_awards	Number of academy awards won	3,464	0.48037	1.115696	0	11
nominations	Number of academy awards nominated	3,464	2.253753	2.269329	1	14
nom_bestpic	Whether the movie was nominated for best picture award	3,464	0.144631	0.351779	0	1
win_bestpic	Whether the movie has won best picture award	3,464	0.025404	0.157372	0	1
nom_bestdirector	Whether the movie was nominated for best director award	3,464	0.113453	0.317191	0	1
win_bestdirector	Whether the movie has won best director award	3,464	0.025693	0.15824	0	1
Movie Connections						
numref	number of previous movies that the current movie references	44,034	4.32618	19.71705	0	1900
numfollow	number of previous movies that the current movie follows as a sequel	44,034	3.075601	12.69171	0	213
numremake	number of previous versions of the movie that the current movie is a remake of	44,034	0.266998	0.799216	0	19

Table 3.6 Summary Statistics

Chapter 4 Results

4.1 Financial Success and Interconnectedness of Movies

In the following table, we present simple correlation results between our innovation index and financial success of a movie, measured by worldwide gross revenue measured in 2010 constant USD. The regression is based on the following model:

$$Gross\ Revenue_i = \beta_1 Num\ Ref_i + \beta_2 Num\ Follow_i + \beta_3 Num\ Remake_i + \alpha_t + \gamma_j$$

Where the subscript i indicates the movie, j indicates the country where the movie was produced, and t the release year.

Table 4.1 shows that movies with more references to earlier works, especially movies that follows earlier ones in a sequel tend to perform better at the box office. This is hardly surprising, though the correlation among different variables might be due to different reasons. For number of references, movies that are viewed by more people would potentially be identified with more connections to earlier ones; and this could explain the positive correlation between gross revenue and number of references. Number of follows and remakes are more objectively measured and should not correlate with number of movie viewers. However, the remaining positive correlation between the two variables and that of gross revenue might suggest the genuine existence of (a negative) relationship between our innovation index and the financial success of a movie. At the bottom line, however, the return rate, defined as gross revenue divided by movie budget was not significant affected by either of the innovation indices. The insignificance suggests that there is no significant inefficiency in the production companies' allocation of funds among different types of movies.

	(1)	(2)	(3)	(7)	(6)
VARIABLES	gross_usd2010	gross_usd2010	gross_usd2010	budget_usd2010	return rate
Num_ref	1.780e+06*** (201,155)	1.897e+06*** (195,326)	1.650e+06*** (195,976)	179,639*** (14,813)	-0.0448 (0.267)
Num_follow	1.349e+07*** (1.996e+06)	9.147e+06*** (1.680e+06)	8.669e+06*** (1.720e+06)	3.633e+06*** (257,510)	-0.615 (2.654)
Num_remake	-7.463e+06 (7.740e+06)	-5.208e+06 (7.467e+06)	-9.014e+06 (7.381e+06)	5.490e+06*** (798,708)	-5.232 (10.53)
Constant	1.887e+08* (1.138e+08)	1.520e+08 (1.000e+08)	1.527e+07 (1.848e+08)	1.248e+07 (1.125e+07)	20.13 (149.6)
Fixed Effect	Year	Year	Year and Country	Year	Year
Sample	movies involving US company in production	movies that has English as one of its official language	All movies	movies involving US company in production	movies involving US company in production
Observations	2,203	2,379	2,546	7,245	2,089
R-squared	0.104	0.102	0.211	0.096	0.010

Table 4.1 Regression Analysis of Financial Success and Movie Interconnectedness

4.2 Effect of Past Academy Awards Nominations on Director's Ability to make Future Movies

In order to trace the impact of award winning on directors, we focus on movies that have less than four directors. Out of the data sample of 966,431 movies, the great majority (962,761, or 99.62%) of it falls into this category.

In the following, we present simple correlation results between several outcome variables and the indicator whether a director has won Best Director Academy Award in year t . The regression is based on the following model:

$$Num\ Movie_{kt} = \beta_1 Nomination_{k,t-1} + \beta_2 Win_{k,t-1} + \beta_3 X_{kt} + \alpha_t + \gamma_j$$

$$Total\ Grosss_{kt} = \beta_1 Nomination_{k,t-1} + \beta_2 Win_{k,t-1} + \beta_3 X_{kt} + \alpha_t + \gamma_j$$

Where the subscript k indicates the director, and t the year, and X_{kt} a bunch of controls for personal characteristics such as age and active years in the industry. α_t and γ_j stand for year and country fixed effect, respectively. Here the coefficient of interest, β_2 , stands for the immediate impact (one-year-lag) of winning a Best Director Academy Award on the director's career.

The following table presents the results. We can see that Academy Award nominations tend to have a bigger impact on the director's career performance than an actual win. A nomination would significantly increase the director's number of feature movies made in the following year, the annual total gross revenue, as well as the average IMDb rating for his movies. Whereas an actual win only significantly affects the gross revenue but not the other two variables.

	(1)	(2)	(4)	(5)	(6)	(7)
VARIABLES	ttnum_movie	ttgross_usd2010	imdb_rating	ttnum_movie	ttgross_usd2010	imdb_rating
win_bestdir	-0.0478 (0.0935)	1.472e+08*** (4.905e+07)	0.181 (0.120)	-0.0404 (0.111)	1.381e+08** (6.773e+07)	0.181** (0.0862)
nom_bestdir	0.178*** (0.0461)	1.594e+08*** (2.946e+07)	0.708*** (0.0597)	0.197*** (0.0551)	1.630e+08*** (4.137e+07)	0.706*** (0.0429)
experience	-0.0210*** (0.000889)	8.020e+06* (4.506e+06)	-0.0191*** (0.00160)	-0.0259*** (0.00776)	-5.916e+06 (1.055e+07)	-0.00205 (0.00635)
Constant	0.806 (0.766)	5.248e+08*** (1.509e+08)	6.793*** (0.210)	-0.949* (0.566)	4.799e+08** (2.015e+08)	7.432*** (0.444)
Fixed Effect	Year and Individual	Year and Individual	Year and Individual	Year and Individual	Year and Individual	Year and Individual
Sample	All directors	All directors	All directors	Directors with Academy Award Nominations	Directors with Academy Award Nominations	Directors with Academy Award Nominations
Observations	475,133	3,340	198,097	3,606	471	3,404
R-squared	0.038	0.108	0.021	0.210	0.273	0.178

Number of id_dirnum	243,417	1,699	94,894	206	117	203
------------------------	---------	-------	--------	-----	-----	-----

Table 4.2 Regression Analysis of Academic awards and various parameters

As a more direct illustration, we compare directors that have won Best Director Academy Awards and those that were nominated in the same year but not win¹.

¹ We only focus on first-time winner of the award and discard individuals when they won for a second time. Moreover, since each winning director is compared to others that were nominated in the same year, the same director might serve multiple times in the control group if he was nominated multiple times.

Chapter 5 Conclusion

Big data analytics has always been of special interest to researchers because it helps to discover hidden patterns, market trends, customer preferences as well as unknown correlations that can exist amongst a large chunk of data, which in turn can help any organization to make informed business decisions about their products in future. With the same aim of providing some value to the billion-dollar movie industry, we conducted data analytics on the movie data set provided by IMDB and Academy Award Association and identified some positive and negative correlation trends that exist in the data which could be of potential benefit to the Film Production houses who are in the business of generating significant revenue by producing films thereby providing a source of entertainment to the audiences worldwide. Following section discusses our findings in detail.

5.1 Significance

We identified that there is a positive correlation between the movie gross revenue and the number of references and number of follows associated with that movie. This signifies that if earning maximum profit is the main criteria for the production house, then they can incorporate more references and follow ups of some of the past movies which were successful at the box office and be somewhat assured that their current movie also would be widely accepted by the audiences based on their previous movie liking trend. Analyzing the same regression from the different angle, may potentially signify that adding innovation to the movie or experimenting with a movie making style may not prove to be beneficial to the production house in terms of generating profit. This could mean that the movie acceptability would be restricted to only a limited set of audiences who are flexible to changes in the movie making style and are more open to accept director's creative work in a positive manner.

The other regression carried out on director's data set signify that receiving an Academy award nomination for best director seems to be of particular advantage to the director's career in the future. As the past data suggest that, there is a positive correlation between award nomination and the number of movies directed by the same director in the future. This could potentially mean that the Oscar nomination fetches the director more credibility in the eyes of production houses and thus s/he receives an increase in the number of films assignments for directing. On the contrary, being a recipient of an Academy award may in turn makes the director more selective in deciding which and how many movies to direct in the future and therefore a negative correlation is observed between the award receipt and the number of future movies directed. The other significant finding is that the Academy award nomination and win both have a positive correlation with the movie gross revenue and therefore it indicates that a movie financial success may be determined in advance by how well establish its director is, in the same field.

5.2 Future Areas of Work

With more and more data being produced on a daily basis, analysis of big data has always been a never-ending process. However, taking into account the current scope of the project and with the existing four csv files generated specifically for this project, we have identified the following areas which can be explored further for providing some more insight into the data.

The criteria for innovation index of the movie can be further refined and the effect of different geographical regions and their economy on the movie financial success can be studied. Secondly, the effect of good movie script and the movie writers' credibility can be analyzed with the movie's financial success. An overall model can be constructed with all the parameters listed, which bear

a positive and significant correlation with the movie gross revenue. Popularity of the Actor/Actress enacting in a movie and their impact on the box office success can as well be analyzed.

References

- [1] Jehoshua Eliashberg, Anita Elberse and Mark A. A. M. Leenders, "The Motion Picture Industry: Critical Issues in Practice, Current Research, and New Research Directions", Published by: INFORMS in Marketing Science, Vol. 25, No. 6, 25th Anniversary Issue (Nov. - Dec., 2006)
- [2] Film industry, Retrieved from https://en.wikipedia.org/wiki/Film_industry
- [3] Theatrical Market Statistics 2016, Retrieved from <http://www.mpa.org/wp-content/uploads/2017/03/2016-Theatrical-Market-Statistics-Report-2.pdf>
- [4] Liran Einav, "Seasonality in the U.S. motion picture industry"
- [5] Michael Luca, "Reviews, Reputation, and Revenue : The Case of Yelp.com"
- [6] IMDB, Retrieved from <https://en.wikipedia.org/wiki/IMDb>
- [7] Bruce W. Herr, Weimao Ke, Elisha Hardy & Katy Börner, "Movies and Actors: Mapping the Internet Movie Database"
- [8] Max Wasserman, Satyam Mukherjee, Konner Scott, Xiao Han T. Zeng, Filippo Radicchi, Luís A. N. Amaral, "Correlations Between User Voting Data, Budget, and Box Office for Films in the Internet Movie Database"
- [9] David D. Galenson and Bruce A. Weinberg, "Creating Modern Art: The Changing Careers of Painters in France from Impressionism to Cubism"
- [10] Max Wasserman, Xiao Han T. Zeng, and Luís A. Nunes Amaral, "Cross-evaluation of metrics to estimate the significance of creative works"
- [11] Wenjing Duan, Bin Gu, Andrew B. Whinston, "Do online reviews matter? — An empirical investigation of panel data", Retrieved from journal homepage - www.elsevier.com/locate/dss
- [12] Gerda Gemser, Mark A. A. M. Leenders, Nachoem M. Wijnberg, "Why Some Awards Are More Effective Signals of Quality Than Others: A Study of Movie Awards"
- [13] Shane Greenstein, Feng Zhu, "Do Experts or Collective Intelligence Write with More Bias? Evidence from Encyclopædia Britannica and Wikipedia"
- [14] David A. Reinstein and Christopher M. Snyder, "The Influence of Expert Reviews on Consumer Demand for Experience Goods: A Case Study of Movie Critics", Published in The Journal of Industrial Economics, Vol. 53, No. 1 (Mar., 2005), pp. 27-51, Retrieved from <http://www.jstor.org/stable/3569753>