THE SECRET RECIPE TO A BLOCKBUSTER HIT

By Anthony Lighterness

ABSTRACT

Increasingly, data science technologies have revolutionized the way that insights inform business. Hollywood is the biggest and oldest of its kind, where the financial success of a blockbuster film may likewise be predictable with linear modelling. The current report aimed to predict gross box office revenues using one general linear model and three variations of generalized linear models (GLMs). For linear prediction, three categorical variables were converted into numeric types, including *Competition*, *MPAARating*, and *OriginalScreePlay*. Next, two other numeric variables, including *GrossBoxOffice*, *EstimatedBudget*, and with non-normal distributions were transformed using a logarithmic (log) function. The best models for predicting *GrossBoxOffice* and its log counterpart were GLM Poisson and GLM normal, respectively. Collectively, this report illustrates statistically significant linear relationships between these two aforementioned response variables and eight others, including *EstimatedBudget*, *Competition Numeric*, *MaxScreenCount*, *Log Estimated Budget*, *Genre_Action*, *StartValue_Cast*, *Genre_Adventure*, and *Year*. Therefore, it is concluded that despite a small sample size of 2,330 cases, a restrictive time period of 10 years, and the simplification of linear models, the financial success of a film may be predicted linearly, and film producers can use this insight as a guide to ensure better outcomes for film.

INTRODUCTION

As the Fourth Industrial Revolution unfolds, led by advances in technologies such as machine learning (ML) and artificial intelligence, data-driven decision making is changing the fundamentals of businesses and industries globally. The film industry, particularly in Hollywood, is no exception to this phenomenon. With an annual net worth of US$41.7 billion in revenue, the global box office in Hollywood is the world's oldest and largest national film industry (McNary, 2019). Gross box office revenue drives this business and is measured by the sum money raised in cinema ticket sales. As such, this measure quantitatively signifies the success of a film.

Predicting a film's financial success, while of utmost importance, is considered a great challenge. Previously, it was widely accepted that, 'Hollywood is the land of hunch and the wild guess' – largely due to the uncertainty associated with product demand and how audiences and critics would receive films in theatre (Litman & Ahn, 1998; Sharda & Delen, 2006). While the fundamentals of statistics and ML principles span decades of existence, it was the recent advances in big data storage technologies that enabled businesses to develop insights for data driven decision making. Gross box office is defined by the ticket sales of a film screening in theatre and as such largely defines the success of a film overall. Historical data can therefore be used to identify key variables that constitute the recipe of a blockbuster hit. Furthermore, future filmmakers may use these indicators as a guide to produce increasingly successful films. Statistical methods can be used to contribute to this reality, for example linear modelling.

In statistics, linear regression is said to be the first type of regression analysis that was studied and applied rigorously (Xin, 2009). It's key goal attempts to establish a linear, causal relationship between a response variable and one or more predictor variables (Xin, 2009; O'Brien, 2007). However, standard linear regression models rely on specific assumptions about the data. Of note, it assumes the response variable to be linearly related to the regression coefficients and predictor variables (Bruce & Bruce, 2017; Efron & Hastie, 2016; O'Brien, 2007). Another assumption is that the data shows constant variance and normality, which can be diagnosed and repaired with distribution plots and transformation functions such as logarithm (log) (Bruce & Bruce, 2017; Efron & Hastie, 2016). Lastly, the predictors must lack perfect multicollinearity. This is theoretically calculated by the Tolerance measure and/or the variance inflation factor (VIF), $T = 1 - R_2$ and VIF = 1/Tolerance, respectively (O'Brien, 2007). A tolerance of <0.10 and/or VIF of >10 is reported to indicate multicollinearity (O'Brien, 2007). The variable of interest, gross box office revenue, is a positive quantity that varies over a large scale. As such, GLMs may provide a better framework for modelling the variable of interest as it is more flexible to arbitrary distributions other than normal (Bruce & Bruce, 2017; Efron & Hastie, 2016).

The current report is organised to first describe the source data and its features. Following this, the methodologies of data pre-processing and linear modelling are detailed before discussing the results. These efforts attempt to investigate the predictive capabilities of a set number of available variables for gross box office revenue. Therefore, the null hypothesis to be tested states there is no statistical difference between box office success and the available variables that describe this population of films. The alternative hypothesis states that there is indeed a statistically significant relationship between gross box office revenue and one or more of these variables.

DATA DESCRIPTION

The current dataset, accompanied by a data dictionary, was sourced from the SAS collection of datasets and data dictionaries (Teradata, 2019; SAS Institute Inc., 2019). It initially presents with a sample size of 2,330, 36 variables and no missing cases. The initial number of categorical variables is 6, with the remaining 30 being numeric in nature. These data items are represented in more detail in Appendix A below. The reliability of the data is measured considering the number of missing cases, number of unexpected outliers present, and any relevant information provided in the data dictionary.

Firstly, the Hollywood dataset is described to have been populated with data sourced from a variety of movie databases which was input using both manual and automatic protocols (Teradata, 2019). As such, it is said that the, "*accuracy of the data set cannot be guaranteed*" (Teradata, 2019). While no missing cases in this dataset are observable, one categorical variable, *MPAARating*, does contain 49 cases (2.1%) of a category that signifies, 'unrated'. This could pose a challenge in interpreting analyses if this report focused specifically on the variable in question. Overall, the dataset is deemed suitable for the purposes of the current report. No known interventions or pre-processing that precede the current study can be reported.

METHODS

To test the hypothesis that *GrossBoxOffice* revenues can be predicted linearly by a selection of variables within the dataset of interest, a number of steps were taken to achieve this goal. Firstly, three categorical variables were converted into numeric data types to suit the quantitative needs of the linear models. The variables, including *Competition*, *MPAARating*, and *OriginalScreePlay*, were taken through a two-step process in SAS VA to achieve true numeric value. To start, the variables were duplicated and modified by selecting 'edit'. In the 'Text' compartment of the editing window, 'If, Else' statements were written to create numeric labels for each category. These scripts are provided in Appendix B below. Following numeric labelling, a 'New Calculated Item' was selected. The variable of interest was then applied within the 'Parse' function, specifying the F12 with 0 decimals for each of the three variables.

Since one key assumption of linear modelling is that of normality, the most critical variables were represented with histograms. Three variables, including *EstimatedBoxOffice*, *GrossBoxOffice*, and *MovieLength* were identified to show skewed distributions. New items were calculated by applying the log function to these variables, producing approximately normal distributions. These results are discussed below along with the exploratory data analysis (EDA) and modeling.

A selection of 35 numeric variables of interest were combined and represented by a correlation matrix, shown in Figure 1.1 below. Following this, distribution histograms and boxplots were applied for further EDA and data representation. The methods conclude with the production of general and generalized linear modelling, of which comparisons were compiled and summarised in Figure 1.2 and Appendix E below.

Firstly, the *GrossBoxOffice* and *Log Gross Box Office* variables were set as the response variables. For each, four models were compared, including a general linear model, and three GLMs with varying distribution settings – normal, poisson, and gamma. While the distribution settings were changed, the remaining default settings were not, allowing simple comparisons considering the increased flexibility to the non-normal distributions of *GrossBoxOffice* and *Log Gross Box Office*.

RESULTS AND DISCUSSION

The descriptive statistics, summarised in Table 1.1 below, show some outliers, such as in *GrossBoxOffice* and *EstimatedBudget*, resulting in significant standard deviations. However, these outliers are not atypical in the Hollywood film industry. Furthermore, Figure 1.2 below indicates that hiding these outliers show appropriate trends in increasing *BoxOfficeClass*. The latter variable describes the success of a film by discretizing the sample population into ten ranges of gross revenue. These boxplots display *EstimatedBudget* grouped by films' *Competition*, and whether or not an original screen play was used, display appropriate increasing trends with film success, as represented by *BoxOfficeClass*. This, together with the fact that the R-square correlation between *GrossBoxOffice* (i.e. the numeric, continuous equivalent of *BoxOfficeClass*), and *EstimatedBudget* is 0.68, is indicative of a slight linear relationship.

The results of the current report show that both *GrossBoxOffice* and its log equivalent may be predictable linearly using a select number of variables. Firstly, log transformations of *GrossBoxOffice, EstimatedBudget,* and *MovieLength* improved the distribution and skewness. For example, Figure 1.1 and Table 1.1 show that the absolute magnitude in skewness decreased from 5.48 to 0.67, for *GrossBoxOffice*. Therefore, the *Log Gross Box Office* variable is more suitable in a general linear model. Although *Log Est Budget* remains have a left skewed distribution, the magnitude of skewness decreased from 3.29 to 1.17. Furthermore, a GLM is sufficiently flexible to accommodate for this.

The correlation matrix identified ten pairs of variables with R-square values greater than 0.60, which are summarized in Appendix D. Figure 1.1 below shows increasing darkness of blue colour that represents the strength of correlations. Interestingly, the log transformation of *GrossBoxOffice* increased this strength for some relationships, which can be seen observing the first two vertical columns from the left of the figure. Of note, its relationship with *MaxScreenCount* increased from 0.527 to 0.870, *StartValue_Cast* from 0.485 to 0.715, and *Log Est Budget* from 0.429 to 0.833, among others. In contrast, the correlation with *EstimatedBudget* decreased from 0.683 to 0.621 when the log transformation was applied to *GrossBoxOffice*. These correlations produce tolerance and variance inflation factor results, also summarised in Table 1.1, which remain within the parameters of no perfect multicollinearity (O'Brien, 2007).

The correlation matrix and boxplot visuals generally support the hypothesis that *GrossBoxOffice*, or its log equivalent, may be linearly related to a selection of these variables. Therefore, eight key variables have been identified as linear predictors when comparing the results of eight linear models. These results are displayed in Figure 1.2 and tabularised in Table 1.1. To predict either *GrossBoxOffice*, or its log equivalent, a general linear model did not produce the best outcome for either. An R-square value of 0.5558 and 0.8547 was produced when *GrossBoxOffice* and *Log Gross Box Office* were the response variables in a general linear model, using all numeric variables excluding *MovieID*. While an R-square value of 0.8547 is considered statistically significant for a general linear model, *Log Gross Box Office* was better predicted in a GLM with normal distribution. In fact, this GLM performed significantly better compared to the Poisson setting, but very similarly to a Gamma distribution.

The AIC produced by the normal GLM, with Poisson, and with Gamma were 6,149.82, 10,981.74, and 6,649.23, respectively. In contrast, when predicting *GrossBoxOffice*, GLM Poisson was selected. While the AIC value is smallest and most preferable for GLM Gamma, followed by GLM normal, the calculated average square error of these models was highest compared to GLM Poisson. Therefore, the GLM Poisson was chosen as the champion model for predicting *GrossBoxOffice*. This could be due to the non-normal distribution of *GrossBoxOffice* that closely follows a Poisson shape. As seen in Figure 1.3 below, the observed averages are nearly perfectly modelled when *Log Gross Box Office* is the response variable, compared to *GrossBoxOffice*. This further validates the log transformation and having a response variable with near normal distribution. The comparisons of these models identified eight key ingredients to linearly predict either *GrossBoxOffice* and/or *Log Gross Box Office*. As tabularized in Appendix E, in order of decreasing importance, these predictor variables include: *EstimatedBudget, Competition Numeric, MaxScreenCount, Log Estimated Budget, Genre_Action, StartValue_Cast, Genre_Adventure*, and *Year*. Interestingly, most of these aforementioned variables were found in common levels of priorities when comparing all eight linear models. All, unsurprisingly, predicted both *GrossBoxOffice* and *Log Gross Box Office* with p-values of <0.00001. For the brevity of the current report, other predictor variables with less significant p-values were excluded in the summarisation of the findings.

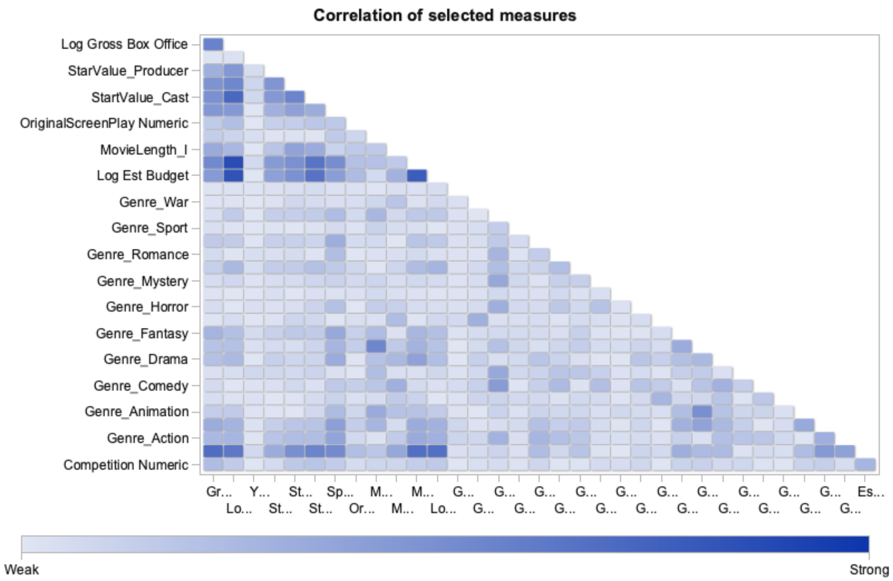

Figure 1.1: Correlation matrix of all numeric variables.

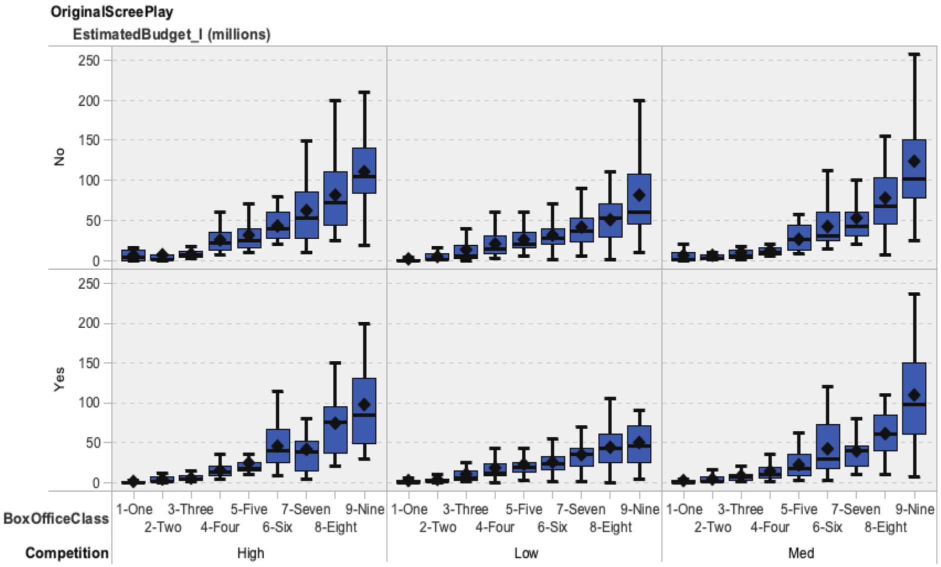

Figure 1.2: Box plots of *EstimatedBudget* (US$ millions) grouped by *OriginalScreePlay* (y-axis) and *Competition* and *BoxOfficeClass* (x-axis).

Data Item	Min	Mean	Median	Max	StDev	Skewness	Correlation R-Square		Tolerance		VIF	
							GrossBox Office	*Log Gross Box Office*	*GrossBo x Office*	*Log Gross Box Office*	*GrossBo x Office*	*Log Gross Box Office*
GrossBoxOffice	50076	6.2x10$_7$	2.2x10$_8$	2x10$_{10}$	1.2x10$_9$	5.48	1	0.564	1	0.436	N/A	2.294
Log Gross Box Office	10.82	16.31	16.91	21.44	2.34	-0.67	0.564	1	0.436	1	2.294	N/A
EstimatedBudget	3000	3.3x10$_7$	2x10$_8$	6x10$_9$	4.2x10$_8$	3.29	0.683	0.621	0.317	0.379	3.154	2.639
Competition Numeric	1	1.52	1	3	0.74	1.04	0.236	0.152	0.764	0.848	1.309	1.179
MaxScreenCount	2	1765.8	2080.5	4468	1300.71	-0.1	0.527	0.87	0.473	0.13	2.114	7.692
MovieLength	45	105.18	102	219	16.63	1.2	0.337	0.262	0.663	0.738	1.508	1.355
MPAARating Numeric	0	3	3	5	0.91	-1.18	-0.134	-0.1	1.134	1.1	0.882	0.909
OriginalScreePlay Numeric	0	1	1	1	0.5	-2	-0.153	-0.22	1.153	1.22	0.867	0.820
StartValue_Cast	1	2.4	2	5	0.91	-0.04	0.485	0.715	0.515	0.285	1.942	3.509
StartValue_Director	1	2.4	2	5	0.73	0.44	0.465	0.536	0.535	0.464	1.869	2.155
StartValue_Producer	1	2.23	2	5	0.69	0.78	0.306	0.447	0.694	0.553	1.441	1.808
SpecialEffects	1	2.4	3	5	0.85	0.12	0.445	0.439	0.555	0.561	1.802	1.783
Log Movie Length	3.81	4.64	4.62	5.39	0.15	0.5	0.309	0.243	0.691	0.757	1.447	1.321
Log Est Budget	8.01	16.24	16.81	20.32	2.01	-1.17	0.429	0.833	0.571	0.167	1.751	5.988

Table 1.1: Descriptive statistics, R-square correlations, and multicollinearity indicators, Tolerance and VIF, of selected numeric variables modelled.

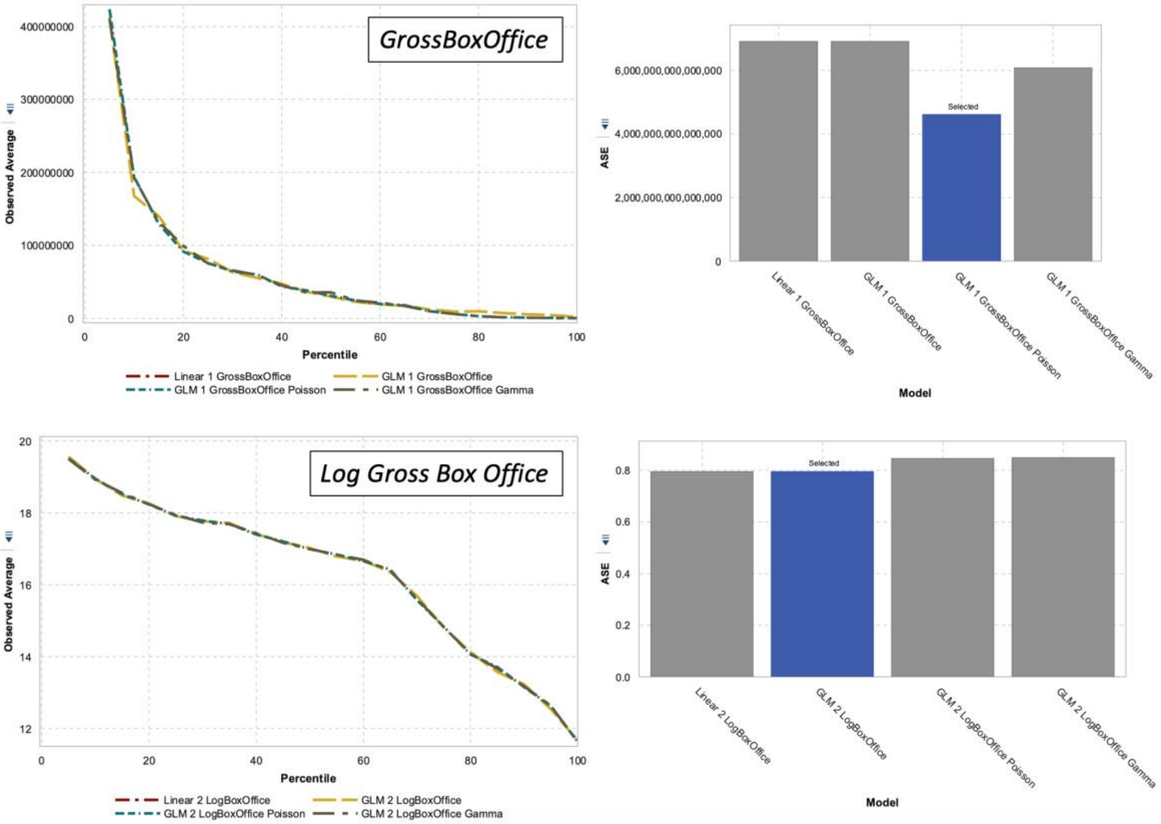

Figure 1.3: Model comparisons for predicting *Log Gross Box Office*.

CONCLUSION

In conclusion, the advent of data science technologies and access to big data has enabled businesses to include data-based evidence for decision making. For big industries, such as the Hollywood film industry, this means better understanding the key ingredients to a blockbuster hit. Linear models have been used widely and most rigorously for good reason. As such, the current investigation sought to use linear models to disprove the null hypothesis that no linear relationships exist between gross box office, or its log equivalent, and a selection of variables provided in a relevant dataset.

The results are in line with the alternative hypothesis that there is some statistical significance between these measurements. Consequently, the current investigation reports eight variables key to linear prediction gross box office, or its log equivalent. Of utmost importance, the estimated budget, competition for the same pool of entertainment, number of theatres the film will be screened at during its debut, and an action genre, most significantly linearly predicts the financial success of a film. For gross box office, a GLM with Poisson distribution best linearly modelled this variable. In contrast, the log transformation of gross box office was best modelled with a normal GLM.

While these results are clear and certain, three key limitations exist in the current investigation. Firstly, linear models are oversimplified and limited to quantitative data. While in this report, qualitative variables were converted to numeric types to accommodate for this, the real world invariably requires a combination of both for optimal accuracy in prediction. Secondly, the quality of the data is highly questionable due to manual and automatic methods used for entry into the database. The reported poor quality by extension cannot by reported, leading to unknown inaccuracy of results. Lastly, the current data only presents 2,330 observations that span one decade of time, from the year 2000 to 2010. Since the release year was identified as the eighth most important variable for linear predictions overall, future research needs to include a wider range of decades to incorporate more data and improve accuracy of the models.

REFERENCING

Bruce, P., & Bruce, A. (2017). *Practical Statistics for Data Scientists*. Sebastopol, California: O'Reilly Media, Inc.

Efron, B., & Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. New York, New York: Cambridge University Press.

Litman, B. R., & Ahn, H. (1998). Predicting financial success of motion pictures. In B. R. Litman (Ed.), The motion picture mega-industry. Boston, MA: Allyn & Bacon Publishing, Inc.

McNary, D. (2019). 2018 Worldwide Box Office Hits Record as Disney Dominates. *Variety*. Retrieved from https://variety.com/2019/film/news/box-office-record-disney-dominates-1203098075/

MPA. (2019). Film Ratings. Retrieved from https://www.motionpictures.org/film-ratings/

O'Brien, R. M. (2007). "A Caution Regarding Rules of Thumb for Variance Inflation Factors". *Quality & Quantity*. **41** (5): 673–690. doi:10.1007/s11135-006-9018-6

Ramesh, S., & Dursun, D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications, 30*, 11. doi:10.1016/j.eswa.2005.07.018

SAS Institute Inc. (2017). *SAS® Visual Analytics 7.4: User's Guide.* Cary, NC: SAS Institute Inc.

Teradata. (2019). SAS Visual Analytics Data Dictionaries. Retrieved from https://www.teradatauniversitynetwork.com/Software/Online/SAS/Student/SAS-Visual-Analytics-Data-Dictionaries/

The output/data analysis for this paper was generated using SAS software. Copyright © [2019] SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

Xin, Y. (2009). Linear Regression Analysis: Theory and Computing. *World Scientific*. Published by World Scientific Publishing Co. Pte. Ltd. in London. ISBN: 9789812834119.

APPENDIX A: TABLE OF ALL DATA ITEMS

Data Item	Description	Data Type	Possible Values
MovieID	Unique identifier.	Continuous, Numeric	10004
GrossBoxOffice	Box-office gross revenue from theatres.		50000
Log Gross Box Office	Log value of 'GrossBoxOffice'.		11.11
MaxScreenCount	Number of screens the movie will be shown at.		11
EstimatedBudget_I	Estimated movie budget.		62054.25
Log Est Budget	Log value of 'EstimatedBudget_I'		8.85
MovieLength_I	Number of minutes the movie runs.		117
Competition	Competition level for same pool of entertainment dollars at debut.	Category	Low, Med, High
MPAARating	Age rating set by Motion Picture Association of America.	Character, Nominal	G, PG, PG13, R, UR, NR, NC17
Competition Numeric	Numeric conversion of 'Competition'.	Discrete, Numeric	Low = 1, Med = 2, High = 3
Genre_*	Content category of a movie, can have more than 1.		0 = no, 1 = yes
MPAARating Numeric	Numeric conversion of 'MPAARating'.		UR = 0, NR = 0, G = 1, PG = 2, PG13 = 3, R = 4, NC17 = 5
SpecialEffects	Signifies the level of technical content and special effects used in the film.		1, 2, 3, 4, 5

StartValue_Cast	Star value of the cast from lowest to highest as per recent box-office success.		
StartValue_Director	Star value of the director involved.		
StartValue_Producer	Star value of the producer.		
YEAR	The year that the movie was shown in theatres.		1-year intervals from 2000 to 2010
OriginalScreePlay	Indicates if movie is based on an original screen play or not.	Binary, Category	Yes, No

OriginalScreePlay 2	Numeric conversion of 'OriginalScreePlay'.	Binary, Numeric	Yes = 1,
			No = 0

Table 1.2: Summarised descriptions of each data item.

APPENDIX B: SCRIPTS FOR TYPE CONVESION

Competition

```
IF ( 'Competition'n Contains 'Low' )
RETURN '1'
ELSE (
  IF ( 'Competition'n Contains 'Med' )
  RETURN '2'
  ELSE (
    IF ( 'Competition'n Contains 'High' )
    RETURN '3'
    ELSE 'Null' ) )
```

MPAARating

```
IF ( 'MPAARating'n Contains 'PG13' )
RETURN '3'
ELSE (
  IF ( 'MPAARating'n Contains 'PG' )
  RETURN '2'
  ELSE (
    IF ( 'MPAARating'n Contains 'G' )
    RETURN '1'
    ELSE (
      IF ( 'MPAARating'n Contains 'UR' )
      RETURN '0'
      ELSE (
IF ( 'MPAARating'n Contains 'NR' )
RETURN '0'
ELSE (
  IF ( 'MPAARating'n Contains 'R' )
  RETURN '4'
  ELSE (
    IF ( 'MPAARating'n Contains 'NC17' )
    RETURN '5'
    ELSE 'Null' ) ) ) ) ) )
```

OriginalScreePlay

```
IF ( 'OriginalScreePlay'n Contains 'Yes' )
RETURN '1'
ELSE (
  IF ( 'OriginalScreePlay'n Contains 'No' )
  RETURN '0'
  ELSE 'Null' )
```

Competition

```
Parse('Competition 2'n, 'F12.')
```

MPAARating

```
Parse('Competition 2'n, 'F12.')
```

OriginalScreePlay

```
Parse('OriginalScreePlay 2'n, 'F12.')
```

APPENDIX C: DISTRIBUTIONS OF SELECT VARIABLES

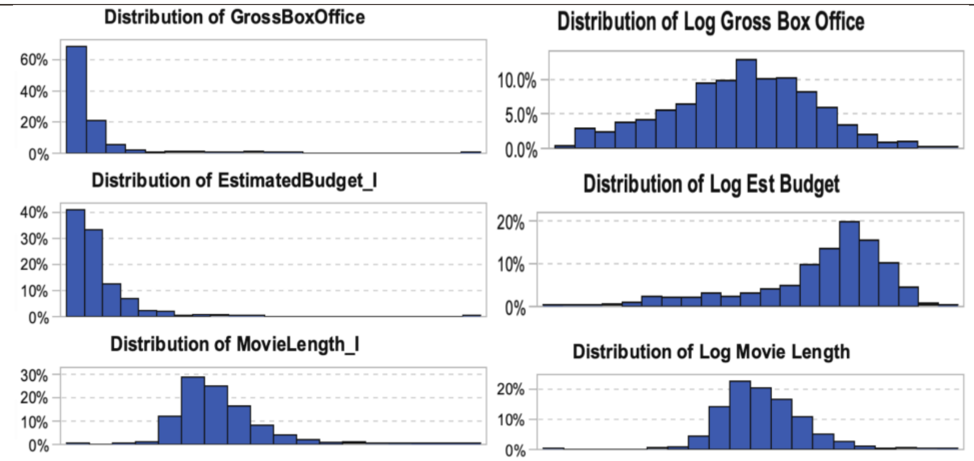

Fig. 1.4: Distributions before and after log conversions of *GrossBoxOffice*, *EstimatedBudget*, and *MovieLength*.

APPENDIX D: CORRELATION RESULTS OF TOP TEN VARIABLE PAIRS

X Axis Variable	Y Axis Variable	Correlation	Tolerance	VIF
MaxScreenCount	Log Gross Box Office	0.8704	0.1296	7.71604938
Log Gross Box Office	Log Estimated Budget	0.8326	0.1674	5.97371565
MaxScreenCount	Log Estimated Budget	0.7761	0.2239	4.46627959
StartValue_Cast	Log Gross Box Office	0.7152	0.2848	3.51123596
GrossBoxOffice	EstimatedBudget_I	0.6829	0.3171	3.15357931
MaxScreenCount	EstimatedBudget_I	0.6709	0.3291	3.03859009
Log Est Budget	EstimatedBudget_I	0.6663	0.3337	2.99670363
StartValue_Cast	Log Est Budget	0.6480	0.352	2.84090909
StartValue_Cast	MaxScreenCount	0.6457	0.3543	2.82246684
Log Gross Box Office	EstimatedBudget_I	0.6207	0.3793	2.63643554
Log Gross Box Office	GrossBoxOffice	0.5637	0.4363	2.29200092
StartValue_Cast	EstimatedBudget_I	0.5537	0.4463	2.24064531

Table 1.3: Correlation, tolerance and VIF results of top 12 pairs of variables.

APPENDIX E: MODEL COMPARISON SUMMARY

Model	R-Square	AIC	F Value	Top 5 Important Variables				
				#1	#2	#3	#4	#5
Linear 1	0.56	87.3x10$_3$	89.9	Estimated Budget	Log Est Budget	Movie Length	Start Value_Cast	Max Screen Count
GLM 1		91.6x10$_3$		Competition Numeric	Genre_Action	Genre_Adventure	Genre_Animation	Genre_Biography
GLM 1 Poisson*		49.1x10$_{12}$		Competition Numeric	Estimated Budget	Genre_Action	Genre_Adventure	Genre_Animation
GLM 1 Gamma		82.3x10$_3$		Estimated Budget	Max Screen Count	Log Est Budget	Start Value_Cast	YEAR
Linear 2	0.85	1,867.56	422.2	Max Screen Count	Log Est Budget	Start Value_Cast	YEAR	Estimated Budget
GLM 2**		6,149.82		Estimated Budget	Max Screen Count	Log Est Budget	Start Value_Cast	YEAR
GLM 2 Poisson		10.9x10$_3$		Estimated Budget	Max Screen Count	Log Est Budget	Start Value_Cast	YEAR
GLM 2 Gamma		6,649.23		Estimated Budget	Max Screen Count	Log Est Budget	Start Value_Cast	YEAR

Table 1.4: Summary of model comparisons. All p-values <0.00001. Linear models that were selected to best model **GrossBoxOffice* and ***Log Gross Box Office*.