Business Statistics Written Report

McGill University

April 17th, 2019

MGCR-271: Section 002 - Team 10

Léo Blanc 260844493

Vincent Copti 260868683

Khalël Fung 260870538

Tracy Li 260785265

Matthew Rugel 260830003

Mattéo Trulli 260841120

Jonathan Zhang 260832478

Introduction

Business statistics is a course meant to educate students of the power of probabilities and learn how to apply statistics in everyday situations. The aim of the assignment is to test material learned from business statistics on real data. Out of the many topics available, we chose to analyze movies because we were curious about what factors make a good movie. We also enjoy watching movies and knew that there would be a large sample for us to choose from. We wanted to discover if there was a correlation between factors. For example, the relationship between genre and budget, duration and genre, duration and budget, or any other variable we could come up with. The goal of the project is to collect a random sample of a population of movies released all around the world. Then we need to formulate three hypotheses to test once we have collected all of our data. Using a five percent significance level, we then verify if we fail to reject the null hypothesis because there is not enough data to prove the contrary. This occurs when the p-value is bigger than the five percent significance level. Or, we reject the null hypothesis for the alternative hypothesis if the p-value is smaller than the five percent significance level. Some other goals of the project would be to run simple regressions and multiple regressions to determine the relations between variables that are most relevant to movies. Chi-square tests will also be another mean of determining the influence of categorical variables for movies.

Statement of hypothesis

To begin, we will present the three hypotheses we decided to test and contribute our reasoning behind our thought process. The first hypothesis we expect is gross profit (dependent variable) will be highest for movies released in January (independent variable). We believed when a movie was released had something to do with the success of the film. We then thought the closer the release date was to the Oscars, the more significance the movie would have. Therefore, generating the most gross profit because of constant marketing. Since the Oscars occur in February we had the impression January would be the biggest month for long awaited movies.

The second hypothesis we expect is gross profit (dependant variable) will increase as IMDB and Rotten Tomatoes ratings (independent variable) increase. Our beliefs were that if a movie was accumulating good reviews from two well known platforms then the movie should be cash flow positive. We also suspected if the ratings were better, than that would translate to the movie in question making a bigger gross profit. Inversely, when movies were gathering mediocre ratings then the movie might have a low positive cash flow to a negative cash flow.

The final hypothesis we expect is gross profit (dependant variable) will be highest for movies that are in the “action” genre (independent variable). The manner in which we reached this conclusion was each member of the team gave their opinion of their favorite movie genre. The result was “action” movie so we were curious to see if “action” film is the best genre in terms of quantity of profits generated per movie genre. An

average gross profit would need to be calculated for each category of genre of film to test our hypothesis.

Data collection process

Next, we will describe the data collection process we went through to get to our sample. We started by deciding that the number of movies we should gather would be a sample of 100 films. Then, we used a random movie generator, Kaggle, to get our sample. Now, we have a 100 movie titles. The following step was to find additional information on each movie. We searched for the main genre of each movie to make calculations more precious. Furthermore, we found the rotten tomatoes and IMDB ratings. Then, we used IMDB.com to collect the budget and gross profit generating in the United-States and Canada. We had to remove a couple of data points in the sample due to missing information needed to complete analysis.

Data description

To describe our data, we focused on the type of variable, the measures of position and dispersion, as well as the relationship between variables and its strength, and finally on the quality of the linear regression.

We collected different variables for each observation. We chose gross profit of the movie as a dependent variable, and release month and year, IMDb and Rotten Tomatoes ratings, genre, budget, and duration of the movie as independent variables. This is 5 quantitative variables (Gross profit, IMDb rating, Rotten tomatoes ratings, budget, duration) and 3 categorical variables (Month, year, genre).

We calculated the measures of position and dispersion, from which we can find some interesting elements (Appendix 1). First of all, we can notice from the mean and median the skewness of the data for different variables. We can see that the distribution of the gross profit is slightly positively skewed (the mean is greater than the median) (Appendix 2). The distribution of the budget is also positively skewed (Appendix 3). Also, the standard deviation for the gross profit is twice larger if we calculate it taking outliers into account (11 outliers for 97 observations), which emphasizes how much these are outlying, and bringing the standard deviation up. Likewise, the standard deviation is twice larger when we calculate it with outliers. Another interesting aspect, is that more than 25% of the data points are showing negative profit (first quartile is negative), and we can calculate that actually 48% of the movies have negative gross profit (we recall that the revenue accounted are only for US and Canada and not worldwide, which explains this point). Moreover, we can notice that the mean and median for the Rotten Tomatoes rating is lower than the IMDb rating, indicating a stricter assessments from users of this website. The distribution of the duration time is peculiar: we can observe 2 spikes in the distribution, indicating 2 common durations, the first around 90 minutes, the second around 115 minutes (Appendix 4). We then calculated the average duration for each genre, and the conclusions were showing the following: each genre has a mean duration of around 90 minutes or around 115 minutes (Appendix 5), that should be the ideal duration for a movie depending on its genre.

To analyze the relationship between our variables, we constructed several scatter plots. We first considered the relationship between gross profit and the ratings (both RT and IMDb). Although both relationships show a positive correlation between gross profit and rating, there are not strong. The linear regressions are positively sloped but relatively flat, and both R square are low: 0.089 for the relationship between gross profit and RT ratings, and 0.152 for gross profit depending on IMDb ratings (Appendix 6 & 7). This means that a movie can be very profitable even if the ratings are relatively negative, and conversely gross profit can be low and ratings positive, even if the trend shows a frail positive correlation. Indeed, IMDb ratings and Rotten Tomatoes ratings are not as strongly correlated as we may think. R square for the best-fit line is equal to 0.565, which shows a certain relationship, but also some major differences between ratings on both websites. We conducted a residual analysis, calculating the residual error. Even if the linear regression is the best way to predict the rating on a website knowing the rating on the other website (in comparison with other non-linear trend lines), the residual error is high, which indicates that the prediction will not be very accurate. For the other linear regressions we made, R square is low, and the residual analysis indicates the weakness of the correlations.

To conclude, we can retain some interesting facts about the position and dispersion of our data, and also some relationships between variables even if most of them are weak.

Methodology

The methodology: multiple regression

1. Test the relationship between duration and gross profit by putting them into a linear regression model. We found that the p-value is 0.82 (Appendix 9), which is too high for us to support the relationship between length of movie and profits.
2. We put IMDB score as new independent variables. We found that the p-value is 0.00069 (Appendix 10), which is less than 5%. Thus, we have enough evidence to support the relationship between IMDB and gross profits. Then, We put Rotten tomato score as new variable. We found that the adjusted r-square decrease to 0. 0967 (Appendix 11).
3. Then, we test the relationship between budget and gross profit.The coefficient is -0.18. The p-value is 0.04 which is lower than 5% (Appendix 12), so we have enough evidence to support the relationship between budget and gross profit. Thus, we put budget as a new variable into our linear regression. We found that the adjusted r-square increase from 0.105 (Appendix 10) to 0.126 (Appendix 13).
4. We put month as the new independent variable, using dummy variables for each month, and leaving February as the one left out, with the dependent variable as Gross Profit, to test our first hypothesis. The adjusted R-square decreases to 0.09 (Appendix 14).
5. We put genre as a new independent variable in our regression model. We choose action as the base case. “Comedy”, “Thriller”, “Drama”, “Romance”,

“Mystery”, “Horror”, “Documentary”, “Sci-Fi”, “Fantasy” are all dummy variables. The adjusted R-square increases to 0.21 (Appendix 15).

Results

1. To test our first hypothesis which is that movies released in January are more profitable than those released in other months, the Ho is that the fact that the movie was released in January or in another month and its profitability are independent variable, while our Ha is that they is a relation between them.

To try to reject Ho we used a Chi-square test with 1 degree of freedom and found a Chi-statistic of 3.28 while the critical value for 1 degree of freedom and a confidence level of 95% is 3.84.

We conclude that we cannot reject Ho and prove that there is a relationship between the fact that the movie was released in January and its profitability.

2. Our second hypothesis is that the higher are the Rotten Tomatoes and IMDB ratings of a movie, the higher will be its profit. To test it we realised a multiple regression with both ratings as independent variables and the profit as dependent variable.

Our Ho is the slope of each regression is 0 and Ha is that at least one slope is different from 0.

The F-stat p-value is 0.007. It’s smaller than our alpha=0.05, thus we can conclude that at least one of the rotten tomatoes and IMDB ratings has influence on the movies profit.

We can then look at each variable individual p-value. The Rotten tomatoes ratings p-value of 0.76>0.05 tells us that there is not enough evidence of a relationship between those ratings and the profit of the movie. However, we can conclude from the IMDB p-value of 0.009<0.05 that the slope of the linear regression is positive (the sample slope is 15614781).

In conclusion, there is no relationship between the Rotten Tomatoes ratings of a movie and its profit but there is a positive relationship between the IMDB ratings and the profit.

3. To test our third hypothesis which is that movies of the action genre are more profitable than movies of other genres, Ho is that the there is no relation between the profitability of a movie and the fact that it is an action movie, while Ha is that those two variables are dependant.

To try to reject Ho we performed a Chi-square test with 1 degree of freedom and found a chi-statistic of 0.87 much smaller than the critical value with 1 degree of freedom and a 95% confidence level which is 3.84.

We conclude that we cannot reject Ho and prove that there is a relationship between the fact that a movie is of the action genre and its profitability.

Recommendation

Firstly, we found that the profits of movies are highly positively related to the rating—IMDB Score and Rotten Tomato Score, which means the public praise is the main factor affecting office box. A movie with a good reputation is more likely to earn a

high profit. Therefore, we recommend movie-makers making high-quality films to satisfy the audiences.

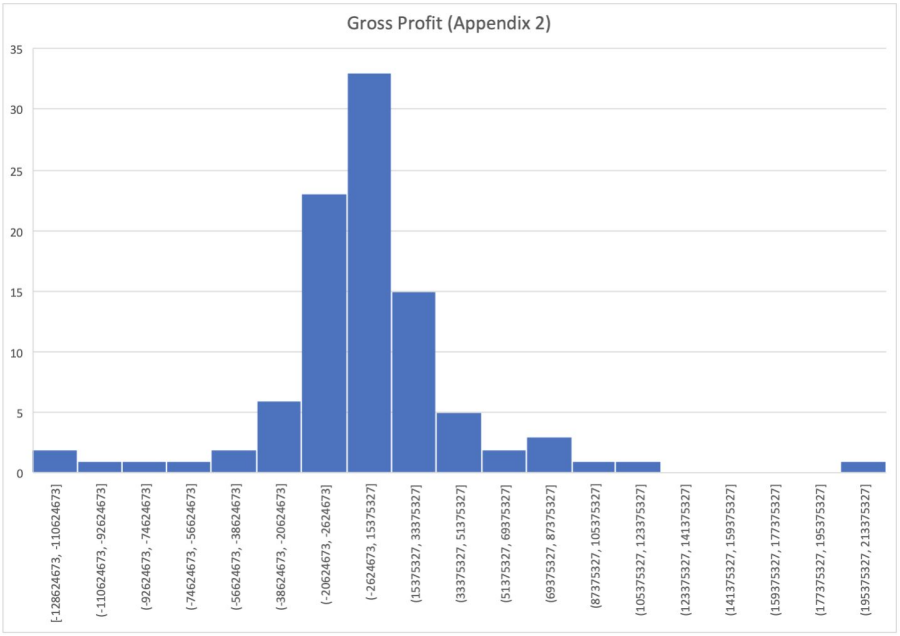
Besides, because of the negative relationship between profit and budget, we recommend the film company not making the high-budget movies which are not likely to achieve a large profits. Last, based on our multiple regression and the hypothesis tests, we do not find the relationship between profit and genre and the relationship between profit and released month. Thus, we cannot give any recommendations about the genres and released month of films, according to our model.

Appendices

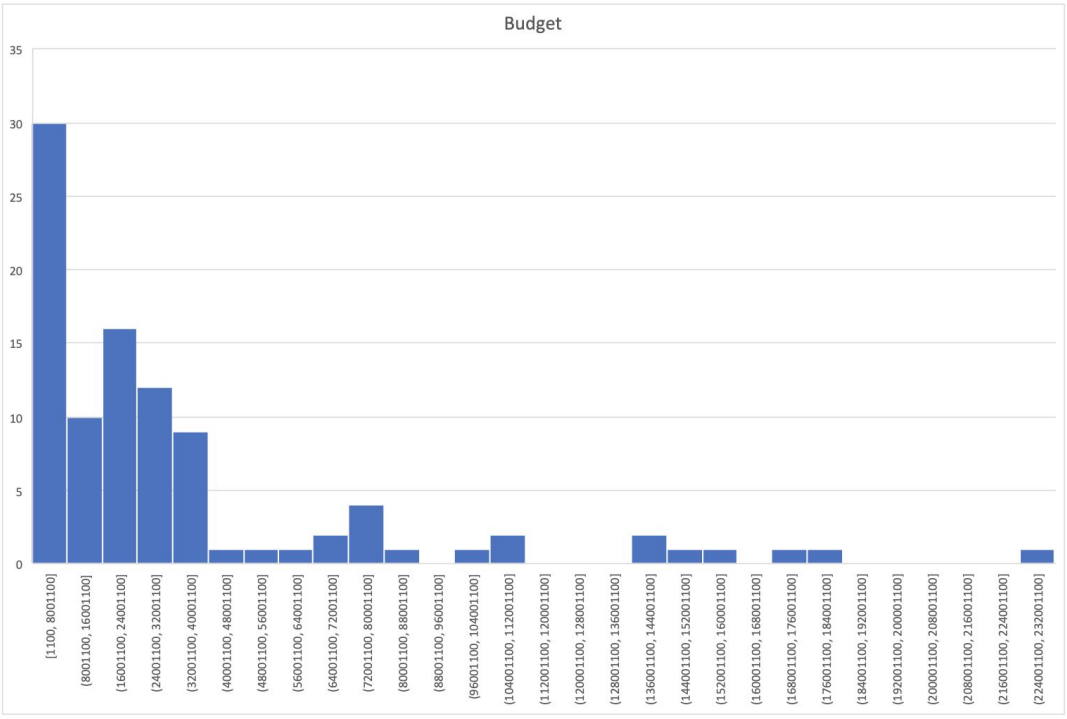
Appendix 1

	Gross Profit	Budget	IMDb Rating	RT Rating	Duration
Mean	5,008,564.39	34,695,513.28	6.42	54.64	109.38
Median	80,016.00	20,000,000.00	6.70	54.00	109.00
Variance	1,764,755,165,579,950.00	2,070,104,953,321,500.00	1.16	668.75	3,048,675,866,405,490.00
Standard Deviation	42,008,989.10	45,498,406.05	1.08	25.86	18.89
Coefficient of Variation	838.74	131.14	16.76	47.33	17.27
Quartile 1	-8,949,272.50	4,100,000.00	5.80	33.50	93.00
Quartile 3	21,918,306.50	40,000,000.00	7.10	76.50	120.50
Interquartile	30,867,579.00	35,900,000.00	1.30	43.00	27.50

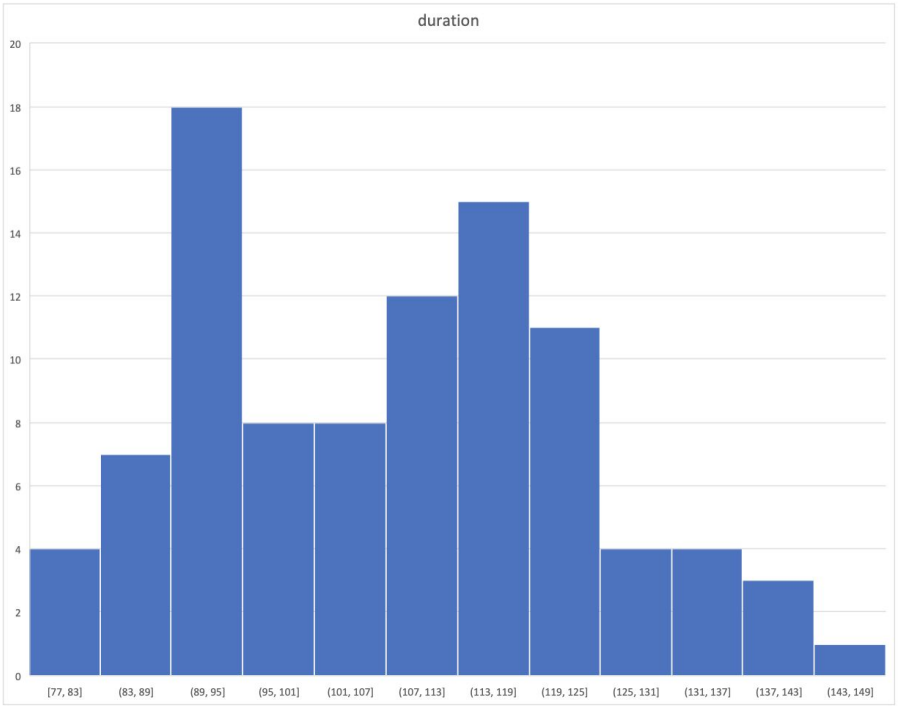
Appendix 2



Appendix 3



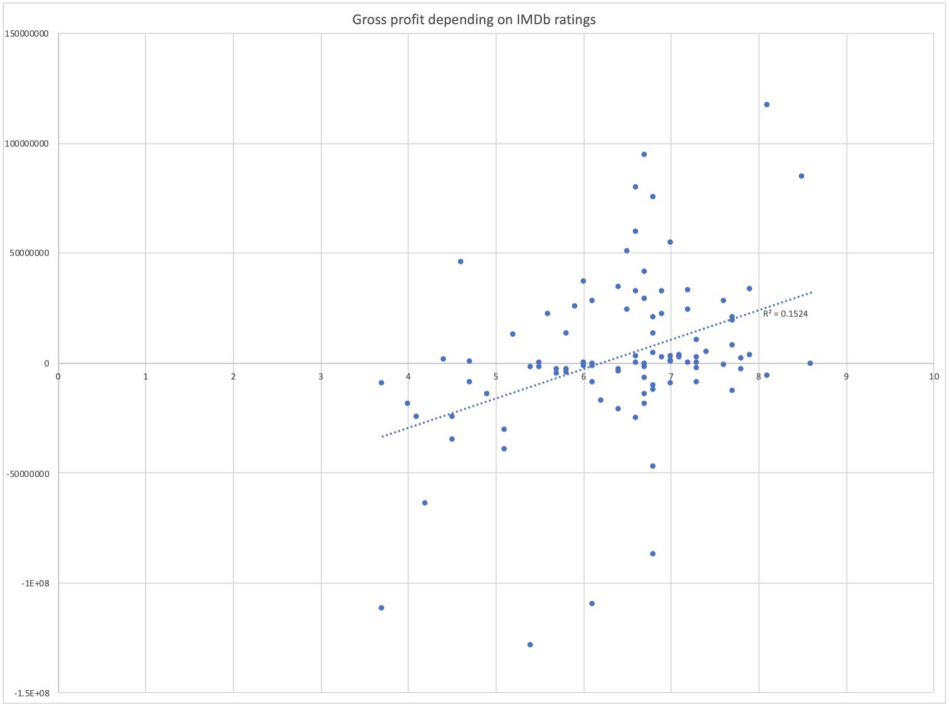
Appendix 4



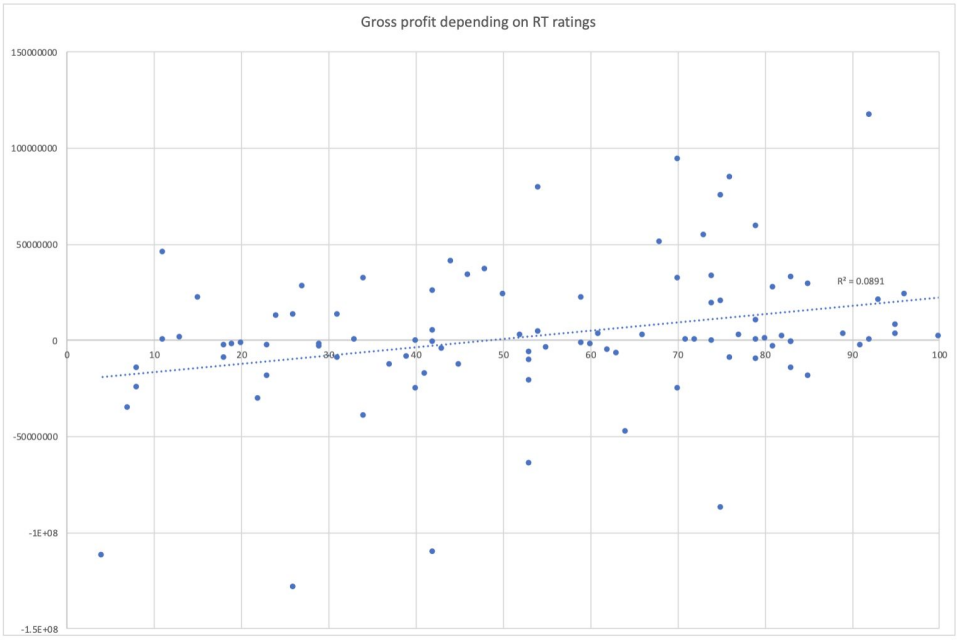
Appendix 5

	Average Duration
Action	114.46
Adventure	99.00
Comedy	96.20
Crime	122.67
Documentary	94.75
Drama	115.00
Fantasy	111.33
Horror	90.50
Mystery	97.50
Romance	101.60
Sci-Fi	116.50
Thriller	107.21

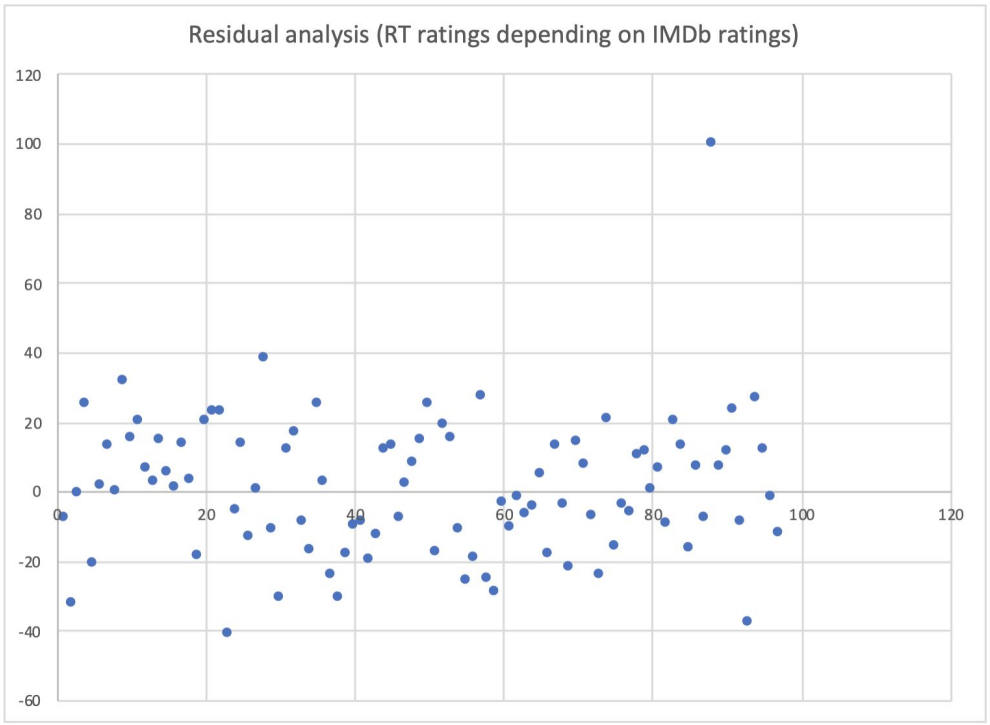
Appendix 6



Appendix 7



Appendix 8



Appendix 9

Regression Statistics	
Multiple R	0.023415
R Square	0.000548
Adjusted R	-0.00997
Standard E	18.98146
Observatic	97

ANOVA

	df	SS	MS	F	ignificance F
Regressior	1	18.77643	18.77643	0.052114	0.819916
Residual	95	34228.11	360.2959		
Total	96	34246.89			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	Upper 95.0%
Intercept	109.3287	1.941096	56.32316	8.63E-75	105.4751	113.1822	105.4751	113.1822
X Variable	1.05E-08	4.61E-08	0.228285	0.819916	-8.1E-08	1.02E-07	-8.1E-08	1.02E-07

Appendix 10

R Square	0.114563
Adjusted R	0.105243
Standard E	1.01777
Observatic	97

ANOVA

	df	SS	MS	F	Significance F
Regression	1	12.73242	12.73242	12.29168	0.000696
Residual	95	98.40634	1.035856		
Total	96	111.1388			

	Coefficient	Standard E	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0	Upper 95.0%
Intercept	6.377152	0.10408	61.27167	3.58E-78	6.170528	6.583777	6.170528	6.583777
X Variable	8.67E-09	2.47E-09	3.50595	0.000696	3.76E-09	1.36E-08	3.76E-09	1.36E-08

Appendix 11

A	B	C	D	E	F	G	H	I
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.340017							
R Square	0.115611							
Adjusted R	0.096794							
Standard E	39922868							
Observatic	97							
ANOVA								
	df	SS	MS	F	ignificance F			
Regressior	2	1.96E+16	9.79E+15	6.144051	0.003106			
Residual	94	1.5E+17	1.59E+15					
Total	96	1.69E+17						
	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	Upper 95.0%
Intercept	-8.5E+07	28657338	-2.95597	0.00394	-1.4E+08	-2.8E+07	-1.4E+08	-2.8E+07
Rotten ton	-79639.5	238617.7	-0.33375	0.739309	-553420	394141.5	-553420	394141.5
imdb_scor	14652067	5735069	2.55482	0.01223	3264954	26039180	3264954	26039180

Appendix 12

Regression Statistics	
Multiple R	0.205748
R Square	0.042332
Adjusted R	0.032251
Standard E	41324702
Observatic	97

ANOVA

	df	SS	MS	F	ignificance F
Regressior	1	7.17E+15	7.17E+15	4.199313	0.043197
Residual	95	1.62E+17	1.71E+15		
Total	96	1.69E+17			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	Upper 95.0%
Intercept	11604553	5286758	2.195023	0.030598	1109012	22100094	1109012	22100094
X Variable	-0.18996	0.0927	-2.04922	0.043197	-0.37399	-0.00593	-0.37399	-0.00593

Appendix 13

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.379999
R Square	0.144399
Adjusted R	0.126195
Standard E	39267719
Observatic	97

ANOVA

	df	SS	MS	F	ignificance F
Regressior	2	2.45E+16	1.22E+16	7.932174	0.000656
Residual	94	1.45E+17	1.54E+15		
Total	96	1.69E+17			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	Upper 95.0%
Intercept	-7E+07	24856688	-2.81271	0.005981	-1.2E+08	-2.1E+07	-1.2E+08	-2.1E+07
budget	-0.16028	0.08853	-1.81051	0.073412	-0.33606	0.015494	-0.33606	0.015494
imdb_scor	12536083	3743605	3.348666	0.00117	5103067	19969099	5103067	19969099

Appendix 14

Regression Statistics	
Multiple R	0.458059
R Square	0.209818
Adjusted R	0.096935
Standard E	39919763
Observatic	97

ANOVA

	df	SS	MS	F	ignificance F
Regressior	12	3.55E+16	2.96E+15	1.85872	0.05156
Residual	84	1.34E+17	1.59E+15		
Total	96	1.69E+17			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	Upper 95.0%
Intercept	-6.4E+07	28104618	-2.26458	0.026114	-1.2E+08	-7756104	-1.2E+08	-7756104
February	-2.3E+07	18235493	-1.28435	0.202551	-6E+07	12842545	-6E+07	12842545
March	-1.8E+07	18302521	-0.95781	0.340906	-5.4E+07	18866205	-5.4E+07	18866205
May	3765179	19523122	0.192857	0.847536	-3.5E+07	42589048	-3.5E+07	42589048
June	-71178.1	15814803	-0.0045	0.99642	-3.2E+07	31378292	-3.2E+07	31378292
July	-7637075	17765209	-0.42989	0.668377	-4.3E+07	27690992	-4.3E+07	27690992
August	-5051286	16719437	-0.30212	0.763307	-3.8E+07	28197149	-3.8E+07	28197149
September	-1.9E+07	16694911	-1.13378	0.260113	-5.2E+07	14271346	-5.2E+07	14271346
October	-1.1E+07	17369272	-0.65157	0.516458	-4.6E+07	23223421	-4.6E+07	23223421
November	478475.9	22328729	0.021429	0.982954	-4.4E+07	44881602	-4.4E+07	44881602
December	14675238	15709853	0.934142	0.352908	-1.7E+07	45916003	-1.7E+07	45916003
budget	-0.20498	0.098026	-2.09111	0.03954	-0.39992	-0.01005	-0.39992	-0.01005
imdb_scor	12530613	4016905	3.11947	0.002483	4542558	20518669	4542558	20518669

Appendix 15

Regression Statistics	
Multiple R	0.555143
R Square	0.308183
Adjusted R	0.218654
Standard E	37132149
Observatic	97

ANOVA

	df	SS	MS	F	ignificance F
Regressior	11	5.22E+16	4.75E+15	3.442267	0.000545
Residual	85	1.17E+17	1.38E+15		
Total	96	1.69E+17			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	Upper 95.0%
Intercept	-1E+08	26065802	-3.82238	0.000251	-1.5E+08	-4.8E+07	-1.5E+08	-4.8E+07
Thriller	14840014	13646496	1.08746	0.279908	-1.2E+07	41972905	-1.2E+07	41972905
Comedy	33472610	15136417	2.211396	0.029692	3377357	63567864	3377357	63567864
Drama	-1.7E+07	11964392	-1.38984	0.168207	-4E+07	7159836	-4E+07	7159836
Romance	12616610	19498513	0.647055	0.51934	-2.6E+07	51384877	-2.6E+07	51384877
Mystery	-7063953	28254727	-0.25001	0.803183	-6.3E+07	49114014	-6.3E+07	49114014
Horror	24335094	28192211	0.863185	0.390465	-3.2E+07	80388763	-3.2E+07	80388763
Document	-1.7E+07	21417794	-0.80661	0.42214	-6E+07	25308408	-6E+07	25308408
Sci-Fi	-2.1E+07	20646333	-1.01899	0.311098	-6.2E+07	20011971	-6.2E+07	20011971
Fantasy	-1.9E+07	18097961	-1.05462	0.294589	-5.5E+07	16897180	-5.5E+07	16897180
imdb_scor	17248191	3933124	4.385366	3.3E-05	9428086	25068296	9428086	25068296
budget	-0.12331	0.092275	-1.33629	0.185022	-0.30677	0.060162	-0.30677	0.060162

Sources of error

While we tried to eliminate any bias in the sample, the sample may be slightly biased as we excluded certain outliers in our calculation, as well as rejected observation with incomplete information (ie. Missing one variable)