**STATISTICAL ANALYSIS OF REVENUE AND TRENDS IN THE MOVIE INDUSTRY**

Aadith Narayan Ravishankar

Naveen Jindal School of Management

Richardson, Texas, USA

aadithnarayan.ravishankar@utdallas.edu

**Abstract**

The movie industry is a billion-dollar industry which has been a consistent provision of entertainment for more than a century. The success potential of a movie is always up to question until after it is released. The reason for this is dependent not on just a single factor but it is dependent on multiple factors. Our point of interest is to find those factors which influence the predictability of a movie being successful in the market. The measure of this success is the Gross revenue generated. This is the exact vision of this research paper, is to figure out the extent of influence which independent variables on the dependent variable Gross. The proposed method looks to employ different statistical analytic techniques like Non parametric T-tests, ANOVA and multivariate and univariate linear regression to accomplish the same. The questions of interest for this project are as follows:

* Whether the audience prefer movies produced with a high budget (>41M$) or with a low budget(<41M$)?
* Does budget for producing movies differ with respect to genre?
* Does the audience prefer a certain genre over other genres?
* Which factors play a positive influence in movies performing well in the market?
* Which is the factor to most consider for the producer to make a profit out of producing a movie?

The dataset will consist of 121 movies from the years 2011 and 2012 with features including Year, Movie title, Runtime, Critic Score, Gross, Budget and an indicator variable for drama movies. Some of the results got were that high budget movies were preferred, there was a difference in the budget investment of producers with respect to genres, the biggest factor to consider whether a movie is profitable or not is the critics score. These results single out particular movies to be successful than the others according to the dataset and provides empirical evidence for each category to deem it profitable or not.

Keywords: ANOVA, Welch T-Test, Linear regression, ggplot2, Backward regression

**Introduction**

America boasts the oldest film industry and produces the largest revenue when compared to any other country. The American film industry produced $43.4 billion in income a year ago, expanding in every one of the previous five years at an annualized pace of only 2.2%, as indicated by a report by the exploration firm IBISWorld. Through 2022, industry revenue will increase at an a rate of 2.0% to $47.9 billion, even though domestic box office is expected to saturate at that time frame, growing at a lackluster annualized rate of 1.1%. Even though the financials are expected to have a detrimental effect, it still appears that the movie industry is perfectly capable of becoming self-sustaining and is still expected to be one of the leading industries in terms of revenue. The billion-dollar industry is worth looking into and investigate in order to find out the catalyst of a booming business. The crux of this industry experiencing a boom is the audience. The audience is so passionate about movies that one does not hesitate in spending money to buy tickets to watch their favorite actors or actresses on screen. This form of chauvinism is integral to the revenue generated in the process. There are more and more producers who are willing to shell out large sums of money to produce a movie with hope and confidence at the same time because they are aware of the ROI they might get if the movie clicks with the audience. So then comes the question, what factors decide if a movie is going to be successful or not? This project looks to address the above question. The aim of this project is not to establish causality because the movies are not randomly generated nor randomly assigned. The dataset used for this project is got from The Data Story and Library (DASL) and Kaggle. There are 3 datasets where using joins and intersections, the final dataset is created. The final dataset contains the following features: Year, Genre, Movie, US Gross, Budget, Runtime and Critic score. The reason for selection of this variables is because on a domain knowledge standpoint, these are the main variables which would determine the quality of a movie and thus its performance in the market if there is a correlation.

**Related Work**

Micheal & Khang [3] discuss predicting the success probability of a movie in its pre-production stage by using Machine Learning and text mining techniques. Here, text mining is used to create mapping between “what” and “who” aspects of a movie, that is, the plot of the movie and the actors respectively and also the “what” and “when” aspect of the movie, the latter being the time of release. Then the project uses Random forest in conjunction with Logit Boost and sentiment analysis to predict the success probability of the movie. The methodology used is strong, but there are two areas in this research which bring doubt to the authenticity of the result. One is that the research is done in the pre-production phase. Agreed that there is a high chance that a movie done by a famous actor and wonderful script is likely to garner the audience’s appreciation. But this is just not enough as there have been many films similar which have failed miserably in the box office. Second, there are much more features that contribute to the success of a film like Runtime, production budget etc. Without the consideration of all the factors, one cannot decide if a movie will be successful or not.

Monalisa & Goutam [4] focuses on using feature extraction using a n-gram model and then applying machine learning classifiers namely SVM, MNB, KNN and ME. Using classifiers gives the polarity, that is, whether a movie succeeds or fails. It does not provide the extent of a movie succeeding or failing in the box office. Also, the feature extraction is done such that, there is no filter on the number of variables taking part in the model. These may give way to redundant variables with correlation which might affect the machine learning model.

Meenakshi et al. [5] focuses on using Naïve Bayes and K means clustering to predict the performance of a movie in the box office. Similar to Monalisa & Goutam [5], the result is a classification output which implies that the extent of success or failure is not provided. The algorithm used the percentage of success of the actor, actress, composer, director, genre to do the classification. The problem with this is, the success of each of the above features are influenced by several other confounding factors. For example, the success of an actress is especially dependent on the genre and the music of the film which would have complemented the actress’ abilities in the movie which would have got her recognition.

Saurabh et al. [6] discusses about movie success prediction in a proper way which the above 3 papers lack. The project focusses on using linear regression and machine learning to predict the extent of success of a movie in the box office. The project involves exploratory data analysis using descriptive statistics assessing for normality and then performing regression and hence the prediction after. However, the project lacks focus on transformation of variables to obtain proper linearity of variables. The linear regression is done by taking advantage of the resistance of the regression properties to heteroskedasticity and kurtosis.

**System Model**

The project uses a dataset of 120 movies which is a result of several joins and intersections from 2 separate datasets [1] [2] which are interlinked by similar movie subsets. The final dataset has the following features:

Year- The year of movie release

Movie- Title of the movie

Genre- The genre of the movie

US Gross- The gross total in millions of dollars in the US box office

Budget- The production budget in millions of dollars

Run Time- The run time or duration of the movie in minutes

Critic Score- The Rotten tomato score of the movie out of 100

A column “profit” is created which is the difference between Budget and Gross in order to assess whether producers get a profit after the release of a movie. A column “budgetf” is created holding Boolean values of “True” if budget is greater than 41M$ and “False” if lesser than 41M$. The dataset is not randomly selected nor is there any room for random assignment. So, causality cannot be established for any of the results obtained in the project, that is the results cannot be extended to other movies which are similar to these movies. But one can come to a conclusion from the analysis the factors to look out for when judging a movie for its performance in the box office. The assumptions in the project are that no two movies are dependent of each other. The assumptions of linearity and constant variance are to be met to perform statistical analysis to get better results and improve accuracy of prediction. This project will assume a significance level of 5%. He packages used in the project will be data.table and ggplot2

Data Exploration

Descriptive statistical analysis

Test for linearity, variance, independence and normality

Parametric or non-parametric statistical analysis

Perform Transformation

Yes

No

Report results

**Fig 1.1: The overall flowchart of the proposed model**

**Problem Statement**

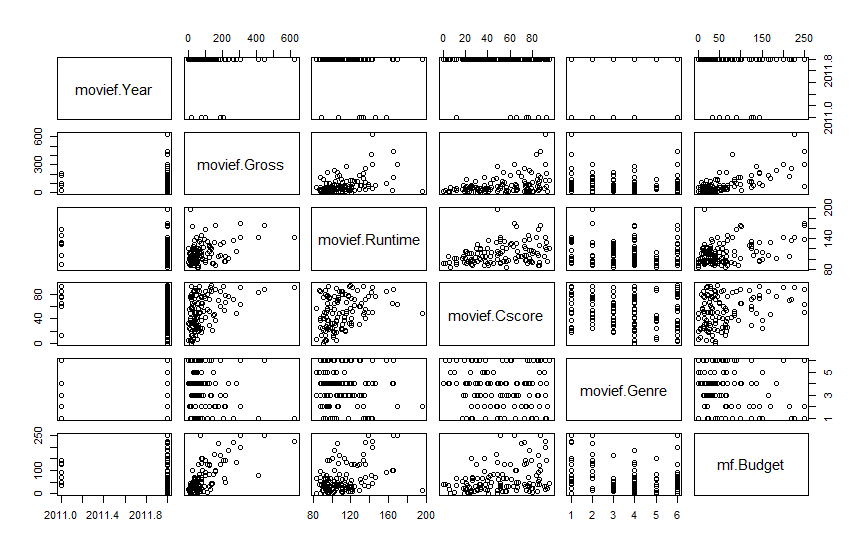
The dataset consists of the features Year, Movie, Genre, Run time, US Gross, Budget and Critic Score. Our point of interest is the US Gross variable which acts as a measure of success of a movie in the box office. The questions of interest are as follows: -

* Whether the audience prefer movies produced with a high budget (>41M$) or with a low budget(<41M$)?
* Does budget for producing movies differ with respect to genre?
* Does the audience prefer a certain genre over other genres?
* Which factors play a positive influence in movies performing well in the market?
* Which is the factor to most consider for the producer to make a profit out of producing a movie?

The intent of this project is to provide direction on where there are differences. The intent is to not to point to a particular entity as the prime attribute over others because trends change over time especially people’s taste in movies.

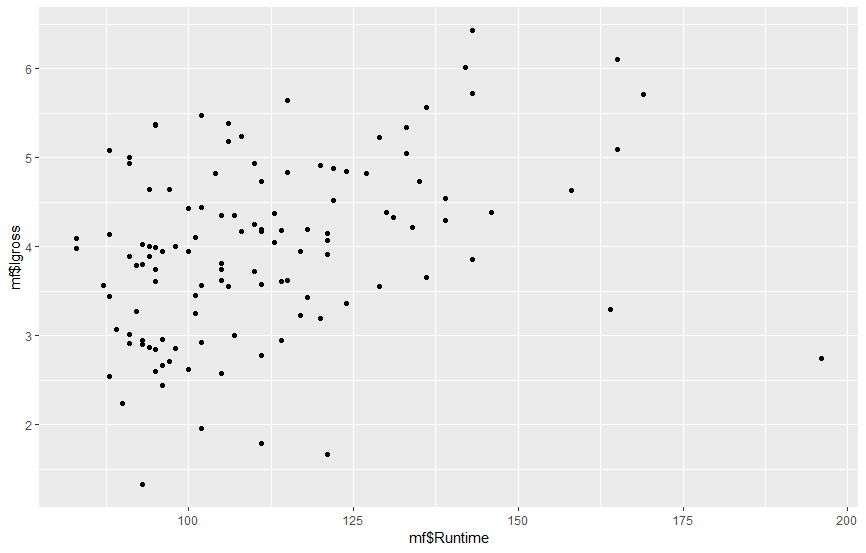
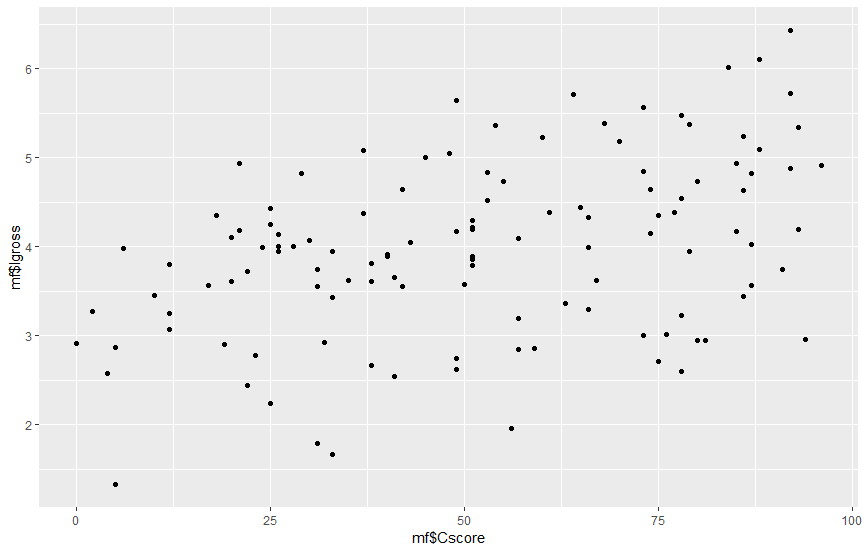
**Solution**

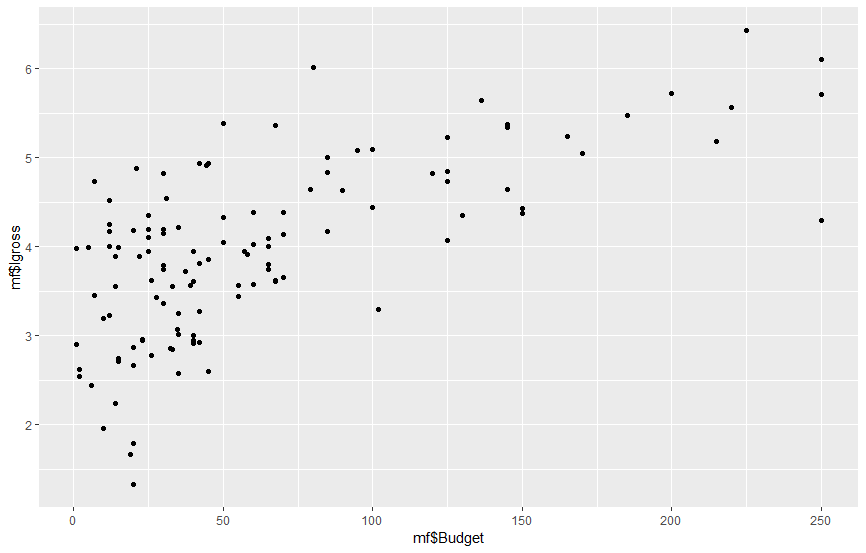
**4.1 Descriptive Analysis**

The first procedure is to understand the data and observe the distribution of the data. This is essential because the assumptions of linearity, constant variance and independence should be satisfied to pursue statistical analysis. The best way to observe the relationship between is by using a matrix scatterplot which provides an overall set of scatterplots for each combination of two variables in the dataset. For analysis, we are excluding the movie attribute because it is not needed for the analysis. 

**Fig 4.1 The matrix scatterplot of the movie dataset**

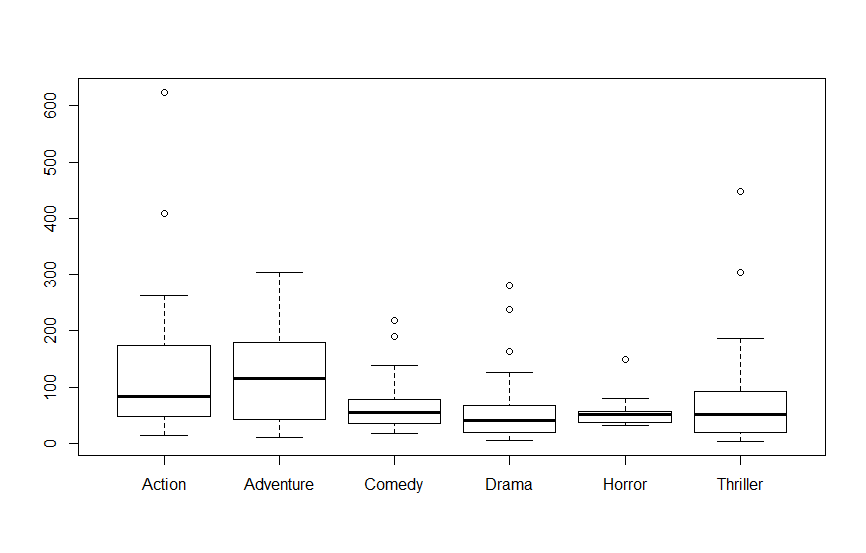
The major point of interest when looking at the scatterplot is “movie.Gross”. There is a positive relationship in most of the combinations involving “Gross” which implies that the mean is straight. But there is increasing variance with increase in x-variable values which indicates heteroskedasticity. Heteroskedasticity is a major detriment when performing linear regression. So, a transformation is needed so that we assess for linearity and constant variance. Generally, for heteroskedasticity, a log transformation is used. So, a feature “lgross” is created which contains all the logged values of “Gross”. Also, another point of interest is the relationship between cscore and budget. The cscore seems to be on a higher level for high budget productions (>41M$). After the transformation, these are the scatterplots with respect to “lgross”:



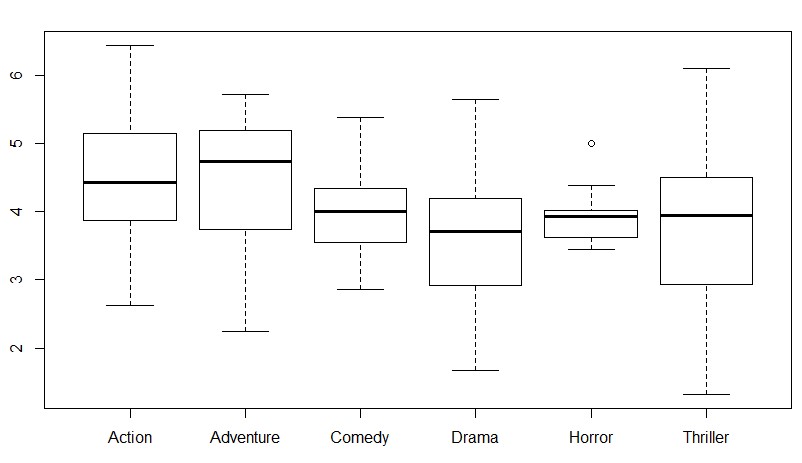


**Fig 4.2 The scatterplots of (a) cscore~lgross (b) runtime~lgross (c) budget~lgross**

As observed, the scatterplots seem to be linear and the heteroskedasticity problem looks to be solved. This should enable us to perform statistical analysis on the data. The next point of interest is to prepare for genre analysis. For that, a boxplot is plotted first with gross as the comparison variable. The intent is to look for skewness and outliers which might affect the analysis and hence make a transformation.

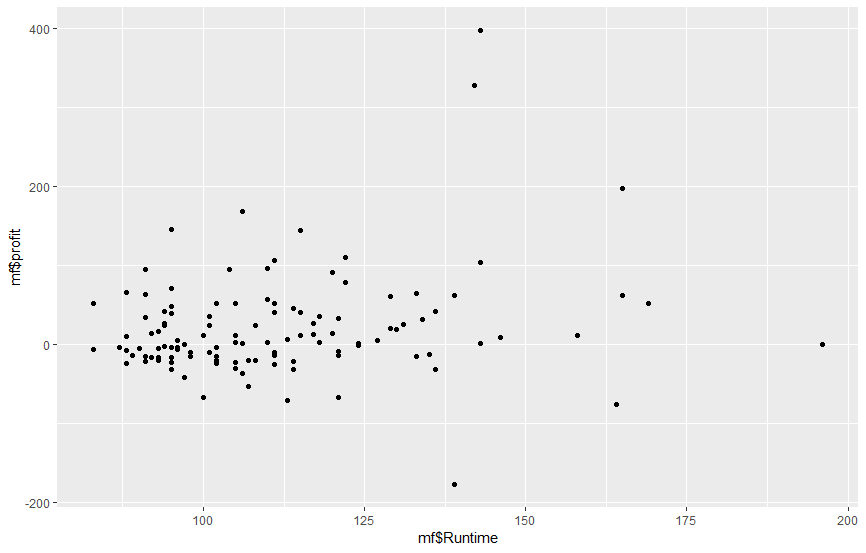
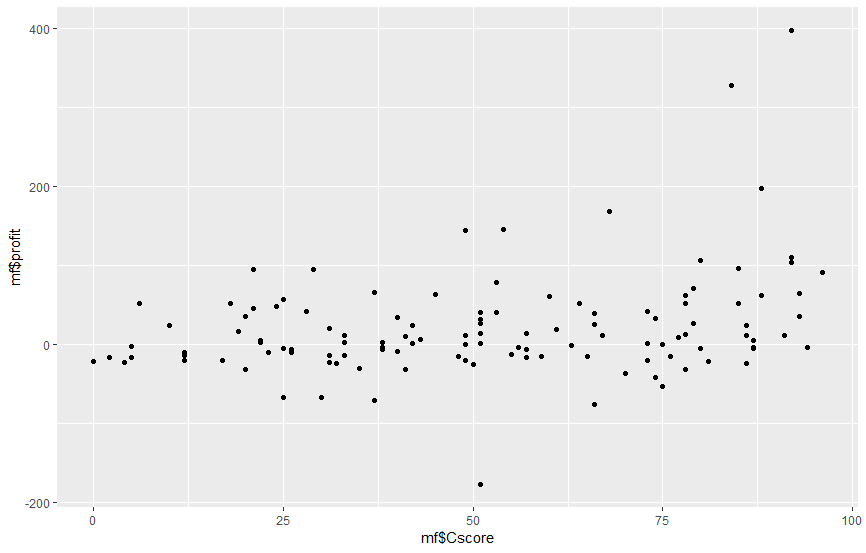
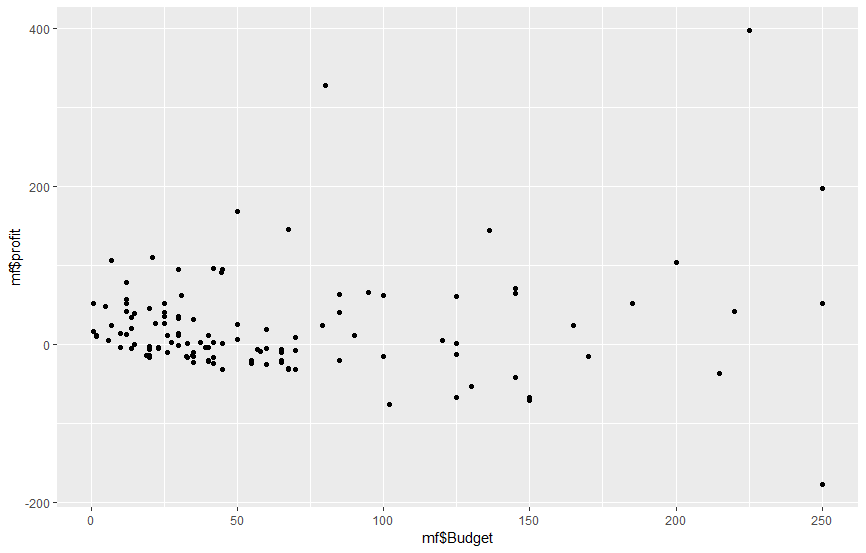


**Fig 4.3 Boxplot between Gross~genre**

As observed, there is a lot of skewness in Action & Horror genres. There is also the presence of many outliers. This does not satisfy the normality assumption for performing ANOVA analysis. There are also multiple outliers which are present which makes it unfit for statistical analysis. To improve the boxplots, a log transformation can be used. The following diagram shows the scatterplot comparing lgross and genre.

**Fig 4.4 Boxplot between lgross~genre**

Now, the boxplots look pretty much normal with almost no outliers. This is fit for statistical analysis. As mentioned before, there is another variable profit which is calculated using the difference between Budget and Gross. The scatterplot of the interactions of profit are as follows:



**Fig 4.5 A scatterplot showing interaction of profit variable with (a) budget (b) cscore (c) runtime**

The scatterplots show some deviations from the assumptions of statistical analysis, and essentially a log transformation is expected. But profits also contain negative values, so a log transformation is out of the picture. Therefore, the analysis is done on profit as it is.

**4.2 Statistical Analysis**

The statistical analysis involves execution of F-test, T-test, ANOVA and linear regression for suitable cases as preferred. This is to find the solution to the questions of interest addressed in the problem statement. Our first question is whether the audience prefer movies produced with a high budget or with a low budget. This is basically a comparison of means of Budget with Cscore which is done over two factors: High and low budget. So what can be deemed as high budget? Movies which have a production budget higher than the median of the total budget are said to be high budget movies. The median of budget was 41M$ found by using the summary statistic of the Budget variable.



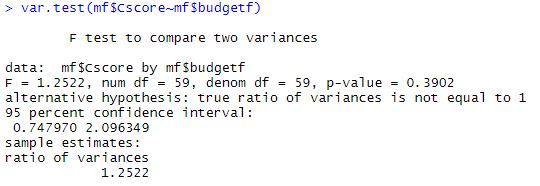
**Fig 4.6 Summary statistics of Budget variable**

Then, a column budgetf is created which stores the Boolean value of whether the budget exceeds 41M$ or not. It is set as “TRUE” if yes, and “FALSE” if not. Then, a F-Test is run comparing the cscore and budgetf to test for equal variances. This is to deem whether to run the T-test with equal or unequal variances.

The null hypothesis for the F-test H0: u1=u2

The alternate hypothesis Ha: u1!=u2

The results of the F-test are as follows:

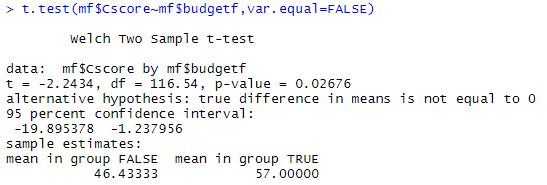


**Fig 4.7 F-test between cscore~budgetf**

The F-test gives a ratio of about 1.25 and we get a p-value of 0.3902 which infers that we reject the null hypothesis and the variances are unequal. So, a Welch T-test would be chosen for this case of unequal variances.

H0: The means are equal

Ha: The means are unequal



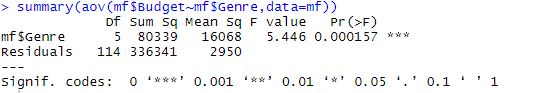
**Fig 4.8 T-test comparing cscore and budgetf**

We get a p-value of 0.026 and hence we reject the null hypothesis and deem that the differences between the means is practically significant. From the means of the two groups, we can conclude that the cscore of high budget films are greater than low budget films. Now since we know that high budget films are preferred by the audience, it would be of great interest to know about the genres which are most preferred and also if the bias in budget is reflected on the bias in genres if present.

There are two questions of interest with respect to genres. One, does budget for producing movies differ with respect to genre? Two, does the audience prefer a certain genre over other genres? For both of these questions an ANOVA test is carried out. This is because there are 6 categories of genres Action, Adventure, Comedy, Drama, Horror and Thriller which have to be compared to continuous valued variables. To answer the first question, an ANOVA test between cscore and genre is carried out.

H0: u1=u2=u3=u4=u5=u6

Ha: At least one of the means is different



**Fig 4.9 ANOVA test comparing budget and genre**

The p-value obtained is 0.00015 which is less than 0.05 and therefore, we reject the null hypothesis. Therefore, we deem that there are certain genres which take more money to produce than other genres. There is a difference in investment for producing movies of certain genres.

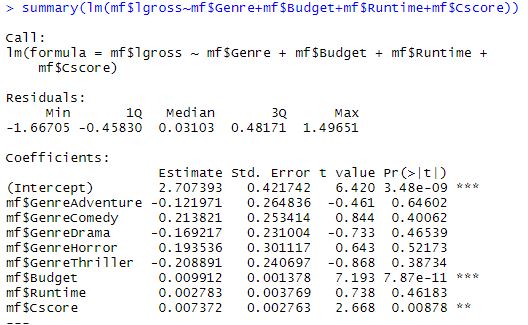
To answer the second question, an ANOVA test is run to compare the means of cscore and genre.



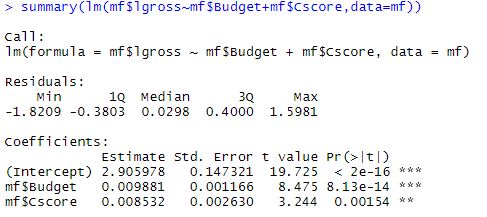
**Fig 4.10 ANOVA comparing cscore and genre**

The p-value is 0.646 which is greater than 0.05 and therefore, we do not reject the null hypothesis. This implies that the audience is unbiased towards all genres when rating the movies. This is quite interesting because there was a difference in means between cscore and budget according to the T-test. But in the ANOVA analysis, there is a difference in the budget allocation for certain genres but the audience are not biased towards the higher budget produced genre. Therefore, there is some kind of an inconsistency in the result which implies that there are some confounding factors at play here.

Now that we have a direction with how the audience perceive movies, the next question of interest is to find out the factors which play a positive influence in movies performing well in the market. For this, a linear regression is run to figure out the influence of independent variables on the dependent variable. In this case, the dependent variable is lgross. The first objective is to select the best combination of independent variables which have a significant effect on lgross. For this purpose, we use a backward regression model and select the best model using a criterion known as the AIC. The backward regression is chosen because there are not too many redundant variables and the best model can be chosen with a few removals of independent variables from the biggest model rather than building up the model from scratch using forward regression.



**Fig 4.11 Summary of the biggest linear model**

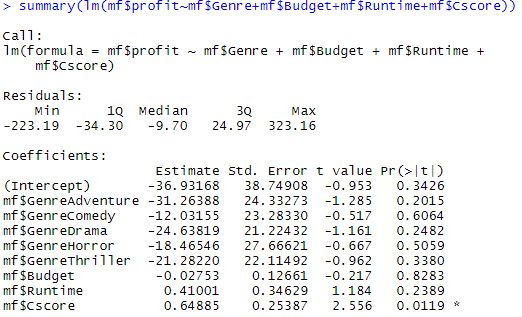


**Fig 4.12 Summary of the best model**

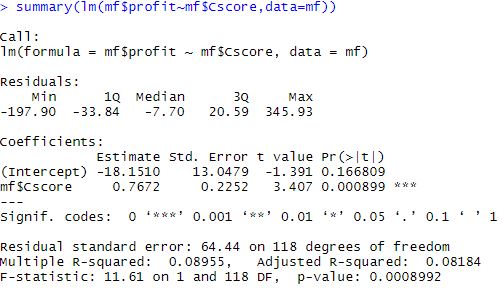
Fig 4.11 shows the summary of the control model which is the model used to start with in backward regression. There are many insignificant variables which do not add value to the dependent variable as denoted by the p-values. The best model got from backward regression as indicated by Fig 4.12 contains just the two variables Budget and cscore. But both the variables have a significant effect on lgross which is the dependent variable. So, this is the best model to explain the change in lgross. The interpretation of the linear model is as follows:

* A 1M$ increase in budget would increase the gross value by 0.99% provided that cscore is kept constant
* A 1-point increase in cscore would lead to a 0.85% increase in gross provided that budget is kept constant

The next question is to find out the factor to most consider for the producer to make a profit out of producing a movie. In order to find this, a profit column is created by finding the difference between gross and budget. A linear model is created with the profit variable as the dependent variable and genre, budget, runtime and cscore as independent variables.



**Fig 4.13 The biggest model for profit**



**Fig 4.14 The best model for profit**

Fig 4.14 provides the best model for profit after backward regression. The only variable which is significant to determine profit is cscore. The interpretation is as follows:

* A 1-point increase in profit leads to a 0.7672 increase in profit.

This result also suggests that the maximum profit a movie can get is 0.7672\*100-18.1510=58M$

The result also suggests that in order to avoid a loss, the movie should get a minimum rating of 24.

**Conclusion**

The findings show that the audience prefers higher budget movies rather than low budget movies. There are certain genres which need to have a higher investment to produce the movie than some genres. But the audience when giving the rating for the movie are unbiased towards genres. This is very interesting as we know that the audience prefer high budget movies, but they are unbiased towards genres which have varying investments to produce. This says that there might be other confounding variables playing a part in the process. The Gross revenue is influenced by two factors, the budget and the cscore. This makes sense because with a higher budget, better looking movies with better actors can be produced which can garner audience support to the movie. The profit, however is influenced only by cscore because generating a profit is a tradeoff between a higher budget movie gaining a higher audience support and low budget movie gaining decent audience support to make a profit out of it. But a low audience support is going to lead to a loss.

**References**

[1] <https://dasl.datadescription.com/datafile/movie-budgets/>

[2] <https://dasl.datadescription.com/datafile/movie-profits/>

[3] Micheal T. Lash and Khang Zhao “Early Predictions of Movie Success: The Who, What and When of Probability” 2016

[4] K. Meenakshi, Maragatham, Neha Agarwal and Ishitha Ghosh. “A data mining technique for Analyzing and predicting the success of Movie” 2018

[5] Saurabh Kumar, Avinay Mehta, Joy Pal. “Movie Success Prediction using Data Mining” 2019

[6] Monalisa Ghosh and Goutam Sanyal. “Performance Assessment of Multiple Classifiers Based on Ensemble Feature Selection Scheme for Sentiment Analysis” 2018

**APPENDIX**

**APPENDIX A**

Transformation of column names to usable keywords

colnames(movief)[colnames(movief)=="ï..Year"] <- "Year"

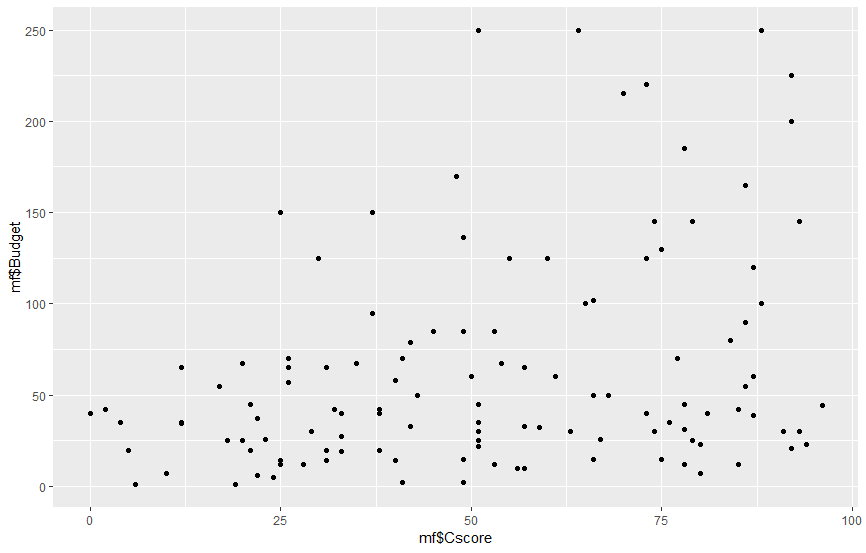
colnames(movief)[colnames(movief)=="US.Gross"] <- "Gross"

colnames(movief)[colnames(movief)=="Run.Time..min."] <- "Runtime"

colnames(movief)[colnames(movief)=="Critic.Score"] <- "Cscore"

**APPENDIX B**

Initial analysis using scatterplot between cscore and budget for the t-test



The scatterplot shows a positive trend as the higher values of budget pertained to higher values of cscore. This provided a foundation for the t-test to be analyzed upon.