DATA 557

Applied Statistics and Experimental Design

ANALYSIS OF POPULARITY AND PROFIT OF MOVIES

Sayani Boral, Kirsteen Ng, Khirod Sahoo, Reeya Pimple

University of Washington

TABLE OF CONTENTS

Abstract

The aim of our analysis is to identify the association between the dependent variables - Profit and Popularity, with the following variables:

- Genre
- Types of Directors
- Types of Actors and Actresses
- Types of Production House

After preliminary data exploratory analysis, we conducted hypothesis testings to confirm the statistical significance of these independent variables. The results from the hypothesis testings show that these factors are statistically significant to the Profit and Popularity of a movie. Finally, we performed multiple linear regression to further investigate the joint association between these variables on the dependent variables.

Introduction

According to Netflix's 2021 Financial Statement[1], Netflix spent around US$13bn on licensing content in 2021 and USD9bn in original content. From the perspective of Netflix, we decided to study which factors would influence the movie's popularity. Meanwhile, Netflix also focuses on generating original content, we are interested in understanding what drives the profit of a movie. We have identified four potential factors from a research paper[2][3], namely Genre, Types of Directors, Types of Actors and Actresses, and Types of Production Houses. The motivation of these variables being chosen will be described in detail under the Statistical Method section.

The following are the questions that we explore in this research.

1. Are all genres equally profitable and popular?
2. Is the popularity and profit between movies produced by top 100 directors the same as non-top 100 directors?
3. Are the popularity and profit between movies casting top 100 actors and actresses the same as movies casting non-top 100 actors and actresses?
4. Are movies produced by a single production house more popular and profitable than movies produced by multiple production houses?

Data Set Description

The dataset is originally from Kaggle, sourced using TMDB API that contains 45,000 movies and 26million user ratings. For each of the questions, we manipulated and generated the following calculated columns.

- Question 1, we tagged the movie to the first genre out of the list of genres given by the original data source.
- Question 2, the boolean column Top 100 Directors was generated by merging the movie dataset with a pre-existing Top 100 Director list from TMDB. If a movie is produced by multiple directors, as long as one of them is in the list above, we will label the movie as 'Yes'.
- Question 3, the boolean columns Top 100 Actors was generated by merging the movie dataset with a pre-existing Top 100 Actors/Actresses list from TMDB. If a movie casts multiple actors, as long as one of them is in the list above, we will label the movie as 'Yes'.
- Question 4, the calculated boolean column of Single House vs Production House was generated by counting the number of Production companies for each movie. If movie is produced by 1 production house it gets tagged as 'Single' and for more than 1 production houses, it will be 'Multi'

Data Set Limitation and Cleaning Process

- The dataset is imbalanced; there are 71% English movies and 29% non-English movies. We decided to focus on English movies for our analysis.
- 85% of movies have been removed due to null values in revenue, budget, or popularity. The final size of the sample data is 4,789.
- Parsed columns with a stringified list of dictionaries to extract names of genres, production companies, actors, and directors.
- The popularity range is not consistent throughout the dataset. There are a few outliers(2.7% of data) where the popularity value lies in the 26-500 range. For the remaining 97.3%, the popularity lies between 0-25.

Statistical Methods

The success of any movie, whether it is released in theaters or on OTT platforms, can be determined by its popularity among its viewers and critics. This metric can be quantified using user ratings, stated as "popularity" score in our dataset, or with the actual money earned by the movie, ie. "revenue". Though it is important to note that all movies are not produced the same, some have a smaller budget and thus a smaller revenue, which does not diminish their success. In this case using profit, calculated by subtracting budget from revenue is a better metric.

Thus, for our analysis, we have used the Popularity and Profit(in USD dollars) of the movie, as the Dependent Variables. The factors which are said to influence a movie as reiterated by the research paper[1], are Genre, Director, Cast, and the Production House. For the purpose of this research, we have delineated the following approach which can be used to answer the four primary questions mentioned above:

A. Stating the Null Hypothesis

B. Performing Descriptive Analysis

C. Choosing a Statistical Test to test the Hypothesis

D. Checking if the Assumptions are met

E. Performing the Statistical Test and analyzing the Significance and Association of Independent Variable to Dependent Variable

The upcoming part of this section will be divided into 2 segments: (i) discussion on assumptions used and methods used to verify assumptions and the (ii) statistical tests used for each question. The first segment is presented in a tabulated format and the statistical tests used are discussed individually for each question mentioned above.

For the first step, we need to verify the underlying assumptions so that we can apply the chosen statistical tests to our sample groups. The verification procedures are similar across all independent variables, therefore we consolidate and summarize the methods and results below.

Assumption Tests and Results

Assumptions	**Result**	**Justification**
Independence of datapoint	Sample data is independent.	Each movie is an individual data point hence independence is assumed to hold.
Normality	Population is normally distributed.	Based on our large sample size(>4k), we adopt the Central Limit Theorem to assume our population is normally distributed.
Constant Variance	● Genre: Popularity and Profit do not have constant variance. Log(Popularity) has constant variance. ● Top Director: Popularity and log(Popularity) do not have constant variance. Profit has a constant variance. ● Top Actors: Popularity, log(Popularity), and Profit do not have constant variance. ● Single/Multi Production House: Popularity, log(Popularity), and Profit do not have constant variance.	We use log transformation on popularity as the scale of popularity ranges widely across all data points. We decide not to remove extreme outliers to maintain the essence of the information in the dataset. Secondly, it helps in some cases(Top directors) in making the variance between the groups constant. We further conducted F-test, Levene's test, and Flingner's test to verify if the variances are constant in log(Popularity) and Profit.
Linearity	The combined model is linear for analyzing both, log(Popularity) and Profit.	A Linear trend is observed while plotting Fitted values vs Residuals for both models.

Table 1

Statistical tests

1. Analyzing the impact of genres on popularity and profit of movies

Description and Motivation

The genre of a movie is a defining factor for its success. Viewers will likely decide whether to watch the movie based on the genre, whether it's young children or thrill-seeking adults. It also decides which award category the movie is eligible for. Thus, it is imperative that we study the effect it will have on defining the movie's success, i.e., the popularity score and profit earned.

There are 18 Genres in our Dataset, namely:

- Action
- Adventure
- Animation
- Comedy
- Crime
- Documentary
- Drama
- Family
- Fantasy
- History
- Horror
- Music
- Mystery
- Romance
- Science Fiction
- Thriller
- War
- Western

Descriptive statistics

In the barplot below, we observe how the Mean Popularity varies across the various Genres, with the Family Genre having a very large Standard Deviation compared to other Genres.

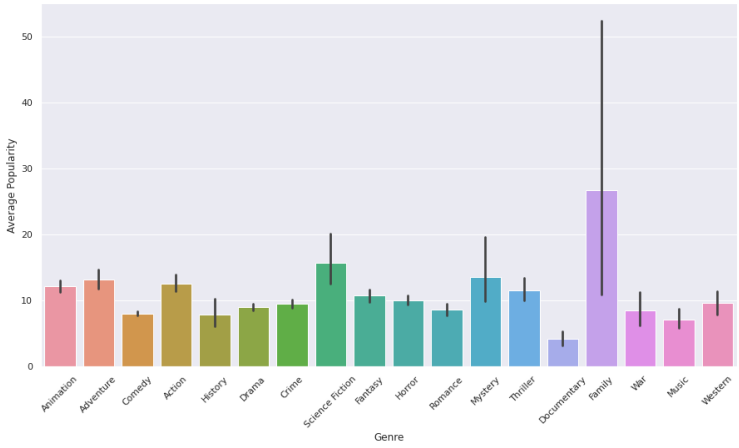

Fig. 1.1 Popularity between different genres

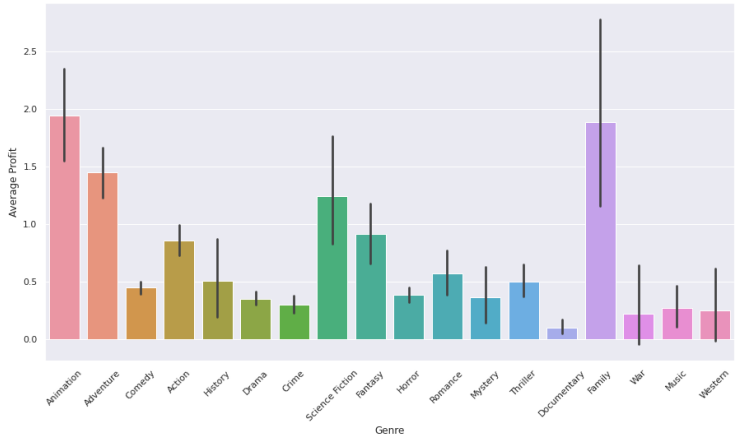

Fig. 1.2 Profit between different genres

Compared to the Mean Popularity, the Mean Profit plotted above displays high variability.

Statistical Tests Performed and results

The various tests performed and results along with the conclusion are tabulated below:

For Popularity/ Profit:

Null Hypothesis: $H_0 : \mu_A = \mu_B = \mu_{C\ldots} = \mu_Q$

Alternative Hypothesis: $H_0 : \mu_A \neq \mu_B \neq \mu_{C\ldots} \neq \mu_Q$

Where μ_{A-Q} : the mean popularity/profit of movies of the 18 main genres represented in the data.

We are presented with 18 Genres and hence perform the ANOVA test to check whether the Mean Popularity is the same across all Genres, but the assumption of Constant Variance is not met. We perform Log transformation on the Mean Popularity which gives us Constant Variance.

The same cannot be done for the Mean Profit variable as we have multiple negative values. Using ANOVA with Non-constant variance can alter the statistical power and the Type I error rate, with the impact being greater when we have more than two groups. Welch' ANOVA[4] uses updated Degrees of Freedom to account for Unequal Variances, where the squared deviation between group means and the general mean are weighted by n_j / s_j^2 instead of n_j.Thus, in this case, we perform Welch's ANOVA, to test the hypothesis for Mean Profit.

The assumptions used here are discussed in Table 1.

Variable	Methods	Test statistic (F-value)	p-value	Significance level	Conclusion
Popularity	**ANOVA test**	10.38	<2e-16	0.05	Rejected H_0
	ANOVA test with log transformation	18.71	<2e-16	0.05	Rejected H_0
Profit(in $ M)	**ANOVA test with Welch correction**	18.81	<2e-16	0.05	Rejected H_0

Table 2: Shows summary of hypothesis tests

Conclusion

As observed in Table 2, we **reject the Null Hypothesis** for both Mean Popularity as well as Mean Profit.

Based on the test results, it can be inferred that the genre is statistically significant towards popularity as well as a profit of the movie.

2. Analyzing the impact of the choice of directors on movie popularity and profit

Description and Motivation

This study[3] claims that the type of directors is one of the classical factors in impacting the popularity and profit of a movie. We also observed that the audience would sometimes consider the director as one of the key factors in producing high-quality movies. Therefore, it is imperative to consider the association of movies with the popularity and profit of the movie.

Descriptive Analysis

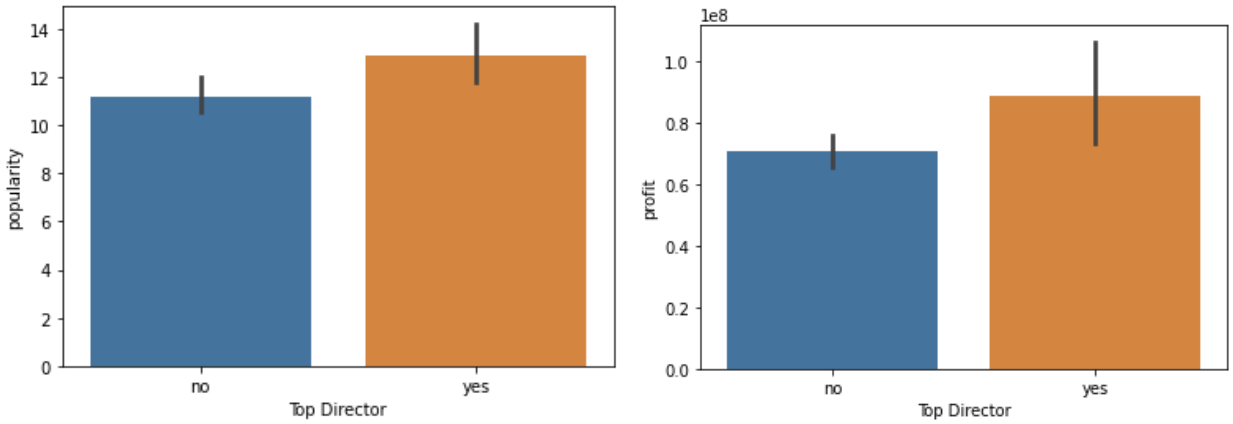

Figure 2.1: Mean popularity and mean profit between the category of Top Director = 'Yes' and Top Director = 'No'.

It can be seen from Figure 2.1 and the table below that despite the number of movies produced by Top Director = 'No' being about 7 times more than the movies produced by Top Director = 'Yes', the mean profit and popularity of the latter group are much higher than the former group. Therefore, we have decided to conduct a one-sided hypothesis testing to verify if Top Director is statistically significant.

The table below is a summary of statistics for our sample data.

	Popularity		**Profit(in $ M)**	
	Top 100 Director	**Non Top 100 Director**	**Top 100 Director**	**Non Top 100 Director**
Sample size	570	4219	570	4219
Mean	**12.91**	**11.2**	**$ 88.88 M**	**$ 70.79 M**
Standard deviation	14.59	23.09	$ 180.89 M	$ 199.87 M
Variance	212.77	532.08	$ 3.994918e+16 M^2	$ 3.272219e+16 M^2

Table 3 Table showing summary of EDA for popularity and profit

Statistical Tests Performed and Result

The following are the null and alternative hypotheses for this category.

For Popularity/ Profit:

Null hypothesis: $H_0 : \mu_{top\ directors} = \mu_{non\ top\ directors}$

Alternate hypothesis: $H_A : \mu_{top\ directors} > \mu_{non\ top\ directors}$

where $\mu_{top\ directors}$: the mean popularity/profit of movies produced by Top 100 Directors, $\mu_{non\ top\ directors}$: the mean popularity/profit of movies produced by non-Top 100 Directors.

Based on the assumptions from Table 1, the following summarizes the hypothesis testing method we chose and their respective results.

Variable	Methods	Test statistics	p-value	Significance level	Confidence Interval	Conclusion
Popularity	**Two sample one-sided Welch test**	2.4218	0.007809	0.05	[0.5468 Inf]	Do not reject H_0
	Two sample one-sided Welch test with log transformation	8.801	<2.2e-16	0.05	[0.0965 Inf]	Rejected H_0
Profit(in $ M)	**Two sample one-sided Z test**	2.3109	0.01044	0.05	[$ 5.36 M Inf]	Rejected H_0

Table 4 Shows summary of hypothesis tests

Conclusion

Between the two tests for popularity, we decide to select the test on popularity with log transformation given there are certain influential data points that we do not wish to remove from the dataset. Therefore, we **reject both the null hypothesis** and conclude that Top Director is a statistically significant variable towards both Popularity and Profit.

3. Analyzing the impact of the choice of actors house on movie popularity and profit

Description and Motivation

This has been a common notion for a long time that a movie casting top actors and actresses tend to be more popular and profitable than other movies. The views have changed in recent decades with changing audiences who crave good content and other factors like screenplay while savoring popcorn. This led us to analyze the effect of casting top actors in a movie. Hence, our null hypothesis is that the mean popularity and profit of movies casting top actors and actresses is the same as other movies.

Descriptive statistics

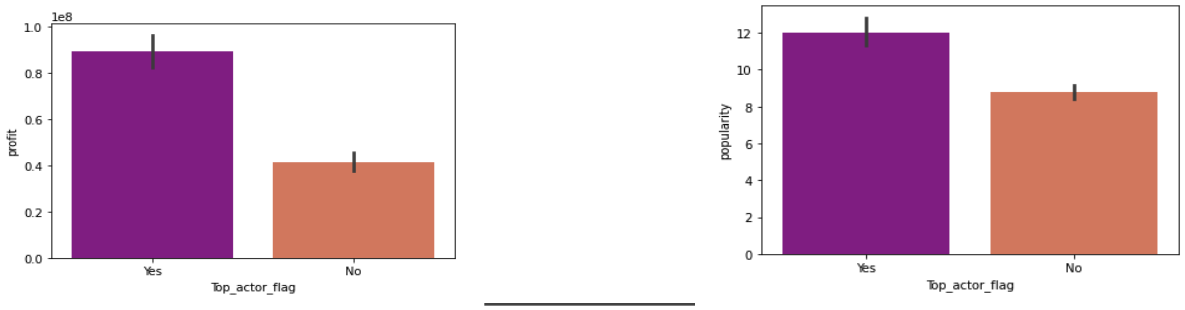

Figure 3.1: Mean popularity and mean profit between the category of Top actor = 'Yes' and Top actor = 'No'.

The movies with Top actors and Non Top actors are almost equally distributed in the sample(Table 5). However, the mean popularity and mean profit of movies with top actors is higher than movies with non top actors(Fig. 3.1)

	Popularity		**Profit(in $ M)**	
	Top 100 Actors/Actresses	**Non-Top 100 Actors/Actresses**	**Top 100 Actors/Actresses**	**Non-Top 100 Actors/Actresses**
Sample size	2299	2490	2299	2490
Mean	**12.00**	**8.79**	**$ 89.30 M**	**$ 41.56 M**
Standard deviation	18.03	9.81	$ 177.51 M	$ 104.70 M
Variance	325.08	96.24	$ 31,509.80 M^2	$ 10,962.09 M^2

Table 5 Table showing summary of EDA for popularity and profit

Statistical tests performed and results

Since constant variance assumption was not met and we had sufficiently large sample size, we decided to use two samples two-sided Z-test. We are using a two-sided test here because we want to explore the effect of top

actors on both sides as it may so happen that despite having top actors, the popularity and profit decrease due to lack of content.

The various tests performed and results along with the conclusion are tabulated below:

For Popularity/ Profit:

Null hypothesis: $H_0 : \mu_{top\ actors/actresses} = \mu_{non\ top\ actors/actresses}$

Alternate hypothesis: $H_A : \mu_{top\ actors/actresses} \neq \mu_{non\ top\ actors/actresses}$

Where $\mu_{top\ actors/actresses}$: the mean popularity/profit of movies casting Top 100 actors/actresses, $\mu_{non\ top\ actors/actresses}$: the mean popularity/profit of movies casting non-Top 100 actors/actresses.

The assumptions used here are discussed in Table 1.

Variable	Methods	Test statistics	p-value	Significance level	Confidence Interval	Conclusion
Popularity	**Two-sample Z test**	7.579	3.493e-14	0.05	[2.384 4.047]	Rejected H_0
	Two-sample Z test with log transformation	15.766	<2.2e-16	0.05	[0.148 0.190]	Rejected H_0
Profit(in $ M)	**Two-sample Z test**	11.217	< 2.2e-16	0.05	[$ 39.40 M $ 56.07 M]	Rejected H_0

Table 6 Shows summary of hypothesis tests

Conclusion

Based on the test results from the above table, it can be inferred that choice of actors does have an impact on popularity as well as a profit of the movie. We also ran 2 sided 2 sample Welch t-test and a robust wald test, and the results from both tests suggested **rejecting the null hypothesis.**

4. Analyzing the impact of the type of production house on movie popularity and profit

Description and Motivation

A movie can be produced and funded by a large single production house like Pixar/MGM/Fox Studios. Sometimes a small production company can collaborate with bigger production houses to produce a movie. We will analyze whether movies that are produced by a single production house or a joint collaboration of multiple production houses have any influence on movie popularity. Also, we will analyze if joint collaboration of production houses makes more profit than single company-produced movies.

Descriptive statistics

We have done some preliminary EDA to understand our sample mean, variance, and group size.

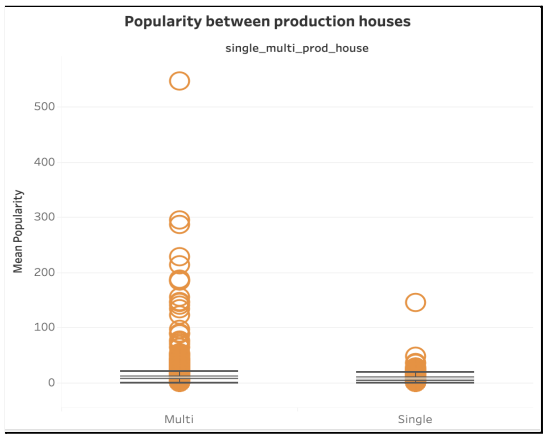

Fig. 4.1 Popularity between production houses

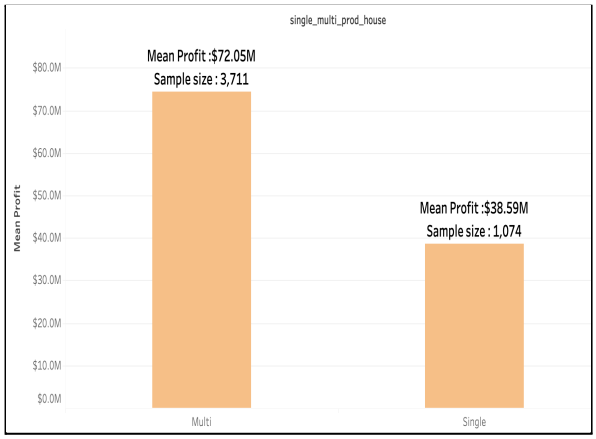

Fig 4.2 Profit between production houses

From Fig 4.1 and Fig 4.2 and Table 7, we can see that the mean profit and popularity for Multi production houses is not the same as that of single production houses.

	Popularity		**Profit(in $ M)**	
	Single house	**Multi House**	**Single House**	**Multi House**
Sample size	1074	3711	1074	3711
Mean	**7.69**	**11.1**	**$ 38.59 M**	**$ 72.04M**
Standard deviation	6.41	15.9	$ 104 M	$ 155 M
Variance	41.17	254.73	$ 10816 M^2	$ 24025 M^2

Table 7 Table showing summary of EDA for popularity and profit

Statistical tests performed and results

From the EDA, we notice that the variance is not constant between the 2 groups of production houses and multi production house popularity and profit is slightly higher than the single production house.

Our null hypothesis is that mean popularity /profit is equal for single and multi-house-produced movies.

Null hypothesis: $H_0 : \mu_{single} = \mu_{multi}$

Alternate hypothesis: $H_A : \mu_{single} \neq \mu_{multi}$

where μ_{single} is the mean popularity/ profit of movies produced by the single production house and μ_{multi} is the mean popularity/ profit of movies produced by multiple production houses

Due to non-constant variance between the groups we decide to use the 2 sided 2 sample Welch test. We use 2 sided test as the mean popularity/profit of movies produced by multi-production may be more than or less than single production houses.

Our assumptions are as stated in Table 1.

Variable	Methods	Test statistics	p-value	Significance level	Confidence Interval	Conclusion
Popularity	2 sample 2 sided Welch test	10.415	2.2E-16	0.05	[2.765, 4.048]	Rejected H_0
	2 sample 2 sided Welch test with log transformation*	11.709	2.2E-16	0.05	[0.129,0.182]	Rejected H_0
Profit(in $ M)	2 sample 2 sided Welch test	8.2098	3.449E-16	0.05	[$ 25.5 M, $ 41.45 M]	Rejected H_0

Table 8 Shows summary of hypothesis tests

Conclusion

As p-value is less than the significance level of 0.05 for all 3 tests, we can **reject the null hypothesis**. Thus we infer that the type of production house (single or multi) does have an impact on the mean popularity/ profit of a movie.

5. Analyzing the associations between all the independent variables and Profit and log(Population)

Description and Motivation

After identifying the statistical significance of individual independent variables, we would like to continue exploring what are the effects of all the independent variables combined on the dependent variables in the study. Therefore, we have built 2 multiple regression models, one each for log(Popularity) and Profit, to answer these 2 questions:

- Understand the strength of association between the independent variables and response variable
- In what way do they(indicated by the magnitude and sign of the beta estimates) impact the dependent variable?

Assumptions

Linearity: We assume that linearity holds in our case as summarized in Table 1.

Multicollinearity: We use the vif() function in the R car package to check for multicollinearity. The table below shows that there is no significant collinearity between the categorical independent variables since the GVIF^(1/(2*DF)) value is around 1. In order to make GVIFs comparable across dimensions, we decide to use the value of GVIF^(1/(2*DF)), where DF is the number of coefficients in the subset[6].

Variable	GVIF^(1/(2*DF))
Genre	1.001671
Top Directors	1.008705
Top Actors	1.001370
Single/Multi Production House	1.028233

Table 9: Multicollinearity using GVIF

Constant Variance: From the previous analysis, we know that log(Popularity) and Profit do not have constant variance in most of our sample groups. Therefore we will use the Robust Wald test function in the R lm package.

Model Results and Interpretation

- **Model 1**: Multiple regression model with log(Popularity) ~ Genre + Top Director + Top Actors + Single/Multi Production House

Null hypothesis: There is no association between the dependent variable (log(Population)) and the independent variables.

Using the Robust Wald Test , we get p-value < 2.2E-16 , so we can reject the null hypothesis.

The following table contains the estimated coefficients, robust t-value, and robust p-value for each variable. We can see that there are variables that have robust p-values less than the significance level of 0.05 and allow us to have sufficient evidence to reject the null hypothesis that all the coefficients of the independent variables are equal.

Multiple Linear Regression Model with Robust Statistics

	Estimate	robust_se	robust_z	robust_p
Main_genre = Adventure	0.01232468	0.02086171	0.59078	0.5547
Main_genre = Animation	**0.07097622**	**0.03136203**	**2.2631255**	**0.0236**
Main_genre = Comedy	**-0.1273139**	**0.01536025**	**-8.2885303**	**0.0000**
Main_genre = Crime	**-0.0907683**	**0.02473824**	**-3.6691495**	**0.0002**
Main_genre = Documentary	**-0.5383292**	**0.09745094**	**-5.5241047**	**0.0000**
Main_genre = Drama	**-0.1541553**	**0.01782682**	**-8.6473848**	**0.0000**
Main_genre = Family	**0.1204866**	**0.05064015**	**2.3792703**	**0.0173**
Main_genre = Fantasy	0.00773073	0.02586398	0.2988994	0.765
Main_genre = Foreign	**-1.0360777**	**0.06448165**	**-16.067791**	**0.0000**
Main_genre = History	**-0.1770181**	**0.06043894**	**-2.9288757**	**0.0034**
Main_genre = Horror	0.0217873	0.0190939	1.1410606	0.2538
Main_genre = Music	**-0.1816124**	**0.07780205**	**-2.3342887**	**0.0196**
Main_genre = Mystery	0.05659576	0.03728267	1.5180179	0.129
Main_genre = Romance	**-0.1657624**	**0.04510906**	**-3.6747022**	**0.0002**
Main_genre = Science Fiction	**0.10690691**	**0.03462512**	**3.0875538**	**0.002**
Main_genre = Thriller	-0.013584	0.02565809	-0.5294233	0.5965
Main_genre = TV Movie	-0.1869891	0.13118181	-1.4254193	0.154
Main_genre = War	**-0.1991273**	**0.06745306**	**-2.9520871**	**0.0032**
Main_genre = Western	-0.0637333	0.03631058	-1.7552275	0.0792
Single_multi_prod_house = Single	**-0.1272007**	**0.01282746**	**-9.916281**	**0.0000**
Top_director = Yes	**0.03138367**	**0.01461519**	**2.1473329**	**0.0318**

	Estimate	robust_se	robust_z	robust_p
Main_genre = Adventure	0.01232468	0.02086171	0.59078	0.5547
Main_genre = Animation	**0.07097622**	**0.03136203**	**2.2631255**	**0.0236**
Top_actor_flag = Yes	**0.16704527**	**0.01115851**	**14.9702186**	**0.0000**

Table 10: Only bolded variables are statistically significant

The following table shows an explicit interpretation of coefficients on the statistically significant variables on Popularity

Variable	Reference Group	Choosing Variable over Reference Group will lead to change in Popularity by
Main_genre = Animation	Genre = Action	7.35% increase
Main_genre = Comedy	Genre = Action	11.95% decrease
Main_genre = Crime	Genre = Action	8.68% decrease
Main_genre = Documentary	Genre = Action	41.63% decrease
Main_genre = Drama	Genre = Action	14.29% decrease
Main_genre = Family	Genre = Action	12.8% increase
Main_genre = Foreign	Genre = Action	64.52% decrease
Main_genre = History	Genre = Action	16.22% decrease
Main_genre = Music	Genre = Action	16.61% decrease
Main_genre = Romance	Genre = Action	15.28% decrease
Main_genre = Science Fiction	Genre = Action	11.28% increase
Main_genre = War	Genre = Action	18.05% decrease
Single_multi_prod_house = Single	Single_multi_prod_house = Multi	12% decrease
Top_director = Yes	Top_director = No	3.18% increase
Top_actor_flag = Yes	Top_actor_flag = No	18.18% increase

Table 11

- **Model 2:** Multiple regression model with Profit ~ Genre + Top Director + Top Actors + Single/Multi Production House

Null Hypothesis: There is no association between the dependent variable (Profit) and the independent variables.

We conducted procedures as Model 1 and obtained a p-value of < 2.2E-16 from the Robust Wald test and we have strong evidence that some of the independent variables have an impact on Profit. The following table contains the regression coefficients, robust z-score, and robust p-value for each variable.

Estimate St	Coefficient	robust_se	robust_z	robust_p
Main_genre = Adventure	**53145613**	**12912061**	**4.1159666**	**0.0000**
Main_genre = Animation	**111361059**	**22387999**	**4.9741408**	**0.0000**
Main_genre = Comedy	**-38253069**	**7111426**	**-5.3790998**	**0.0000**
Main_genre = Crime	**-61928596**	**7853254**	**-7.8857242**	**0.0000**
Main_genre = Documentary	**-58782466**	**7797268**	**-7.5388546**	**0.0000**
Main_genre = Drama	**-54163055**	**7290284**	**-7.429485**	**0.0000**
Main_genre = Family	**103000629**	**41966953**	**2.4543271**	**0.0141**
Main_genre = Fantasy	5931817	14758499	0.4019255	0.6877
Main_genre = Foreign	**-50763081**	**12208786**	**-4.1579137**	**0.0000**
Main_genre = History	**-43178570**	**17520872**	**-2.4644076**	**0.0137**
Main_genre = Horror	**-35113421**	**7257386**	**-4.8383011**	**0.0000**
Main_genre = Music	**-43897528**	**11917069**	**-3.6835843**	**0.0002**
Main_genre = Mystery	**-46958596**	**15169496**	**-3.0955938**	**0.002**
Main_genre = Romance	**-33184502**	**12287111**	**-2.7007571**	**0.0069**
Main_genre = Science Fiction	43232288	23848805	1.8127654	0.0699
Main_genre = Thriller	**-40481332**	**9633695**	**-4.2020566**	**0.0000**
Main_genre = TV Movie	-31415791	38406170	-0.8179881	0.4134
Main_genre = War	**-74808945**	**20946583**	**-3.5714151**	**0.0004**
Main_genre = Western	**-74140776**	**18023370**	**-4.1135912**	**0.0000**
Single_multi_prod_house=Single	**-25181271**	**3872666**	**-6.5023091**	**0.0000**
Top_director = Yes	2557491	6600833	0.3874496	0.6984
Top_actor_flag = Yes	**47506181**	**4186527**	**11.3473964**	**0.0000**

Table 12

The following table shows an explicit interpretation of coefficients on the statistically significant variables on Profit.

Variables	Reference Group	Choosing Variable in col 1 over reference group will lead to change in profit by
Main_genre = Adventure	Main_genre = Action	$ 53 M increase
Main_genre = Animation	Main_genre = Action	$ 111 M increase
Main_genre = Comedy	Main_genre = Action	$ 38 M decrease
Main_genre = Crime	Main_genre = Action	$ 61 M decrease
Main_genre = Documentary	Main_genre = Action	$ 58 M decrease
Main_genre = Drama	Main_genre = Action	$ 54 M decrease
Main_genre = Family	Main_genre = Action	$ 103 M increase
Main_genre = Foreign	Main_genre = Action	$ 50 M decrease
Main_genre = History	Main_genre = Action	$ 43 M decrease
Main_genre = Horror	Main_genre = Action	$ 35 M decrease
Main_genre = Mystery	Main_genre = Action	$ 46 M decrease
Main_genre = Music	Main_genre = Action	$ 43 M decrease
Main_genre = Romance	Main_genre = Action	$ 33 M decrease
Main_genre = Thriller	Main_genre = Action	$ 40 M decrease
Main_genre = War	Main_genre = Action	$ 74 M decrease
Main_genre = Western	Main_genre = Action	$ 74 M decrease
Single_multi_prod_house = Single	Single_multi_prod_house = Multi	$ 25 M decrease
Top_actor_flag = Yes	Top_actor_flag = No	$ 47 M increase

Table 13

Final Conclusion

In our individual analysis we measured the significance of variables in question in impacting the popularity and profit of movies. For making a final conclusion, we used log(popularity) instead of popularity since in one of the questions, log(popularity) helped in holding the constant variance assumption in sample groups. Hence, to be consistent in our final interpretation, we used log(popularity) as our dependent variable.

We observed that performing individual statistical tests for Genre, Top Director, Top Actor, and Type of Production House gives us a result that is statistically significant towards log(popularity), and we obtained a similar result when we combined the features. Performing Multi-Linear Regression using Robust Errors on log(Popularity), we observed that only a few genres: Science fiction, fantasy, horror, mystery are not statistically significant.

Similarly, we observed statistically significant results for all 4 variables, when tested individually towards Profit. While performing Multi-Linear Regression using Profit as the dependent variable, the result suggested that there might be some confounding effect on the variable Top Directors, as its p-value is greater than 0.05 in the combined analysis. This implies that when it is combined with other factors, this variable might be impacted by other confounding variables in the model. This leads us to wonder if there are confounding variables in the dataset that we need to further investigate.

Future Work

We realize that there could be other confounding variables and we suspect budget could be one such variable as the higher budget may attract top actors, top directors, and may bring joint house collaboration. In our future work, we plan to use the budget in our linear regression model and re-visit the parameters and significance in the model. Also, there could be other important features like the runtime of the movie, the ratio of female over male casts that may impact profit and popularity. We plan to explore those as well.

Reference links

[1]Netflix:Netflix Slows Spending on Licensed Content, Focuses on Original Programming | Barron's

[2]Financials of Netflix:
https://s22.q4cdn.com/959853165/files/doc_financials/2021/q4/da27d24b-9358-4b5c-a424-6da061d91836.pdf

[3] Factors Affecting the Success of Movies- A Case Study of Twin Movies

[4] Welch correction for ANOVA : https://www.rips-irsp.com/articles/10.5334/irsp.198/

[5] Interpreting Regression models with Log Transformed variables: FAQ How do I interpret a regression model when some variables are log transformed? (ucla.edu)

[6] GVIF explanation: https://www.tandfonline.com/doi/abs/10.1080/01621459.1992.10475190#.U2jkTFdMzTo

Appendix

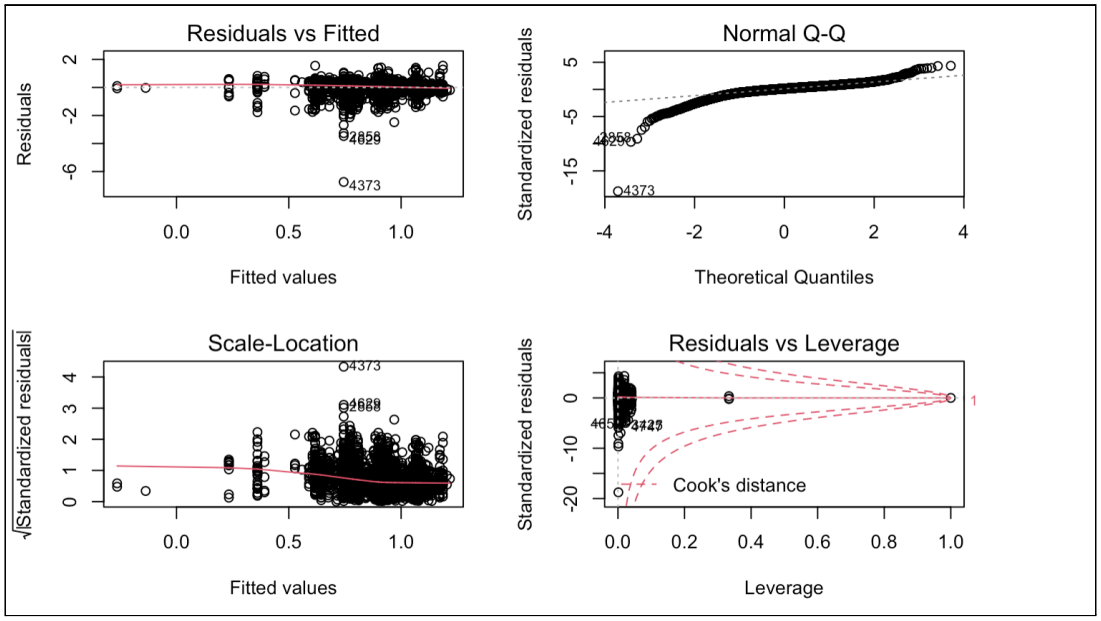

Fig Appendix 1. Linear regression diagnostic plots for log(popularity)

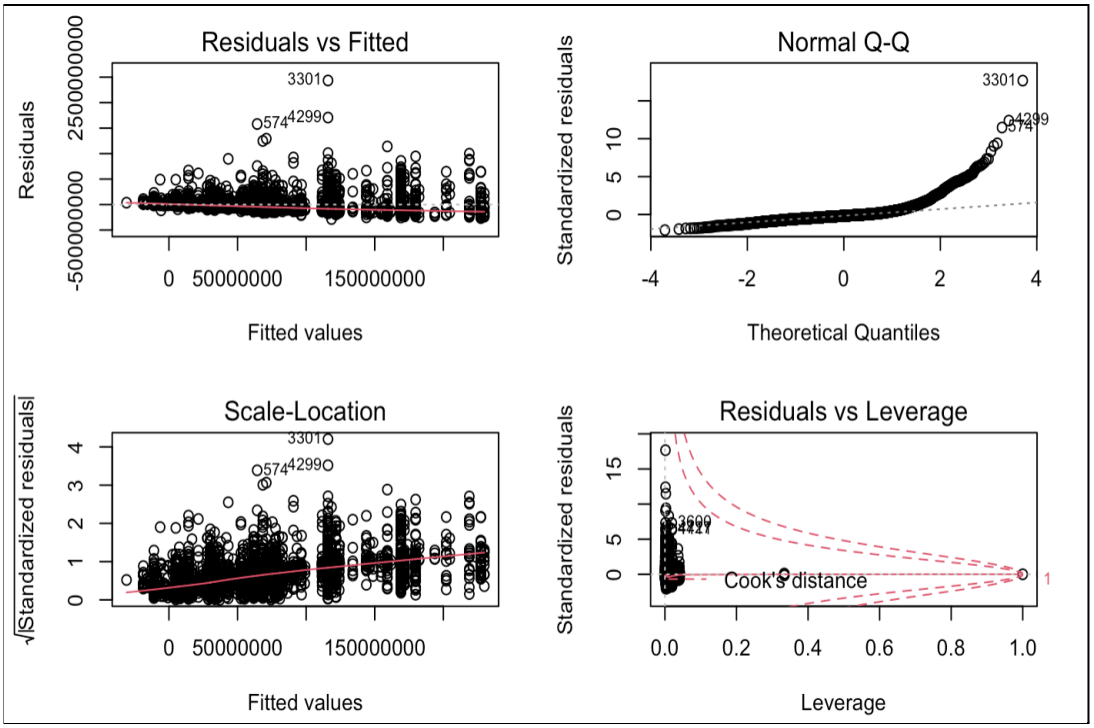

Fig Appendix 2. Linear regression diagnostic plots for profit