SMU
SINGAPORE MANAGEMENT
UNIVERSITY

School of Computing and Information Systems

ISSS616 Applied Statistics Analysis with R – G1
Group Project Final Report

Netflix Data Insights

Group 3:
Aishwarya Sanjay Maloo
Arvind Murali
Le Duy Hoang
Prachi Rajendra Ashani

Table of Contents

1. Introduction

Netflix is one of the most popular media and video streaming platforms in the world – available in 190 countries[1] with over 222[2] million subscribers globally. The platform caters to different markets with curated content tailored for different audiences.

In an extremely competitive landscape of the entertainment industry in which multiple platforms such as Disney+, Amazon Prime, and HBO Max are vying for the same set customers, it is essential for any content streaming platform to gather the right information and insights to retain their market share and accelerate growth in the cut-throat business.

Furthermore, given how transient consumer preferences have become with respect to consuming content, it is imperative for actors, and directors to choose scripts that can help them stay relevant for longer in the industry and for the movie producers to maximize their returns on investment.

2. Objective

The objective of this project is to gain insights on the factors that affect the success of movies to better understand the viewer preferences on Netflix through statistical analysis. The analysis will help us gain insights into the elements that can contribute to a movie's potential box office success and awards. This, in turn, can guide Netflix and movie producers to divert resources to projects that maximize chances of viewer ratings and box office success. Simultaneously, such insights can help actors and directors to understand what movies resonate with their viewers and can make professional decisions accordingly.

For viewers, the information presented in this report can guide their decision making when picking a movie and better understand certain aspects such as genres, actors, and directors that can entice the audience and could potentially make a movie a hit.

Additionally, with the right insights, personalized marketing becomes much more effective, and the platform can suggest a curated list of movies based on consumer preferences. This will drive increased membership rates along with higher customer satisfaction.

3. Data

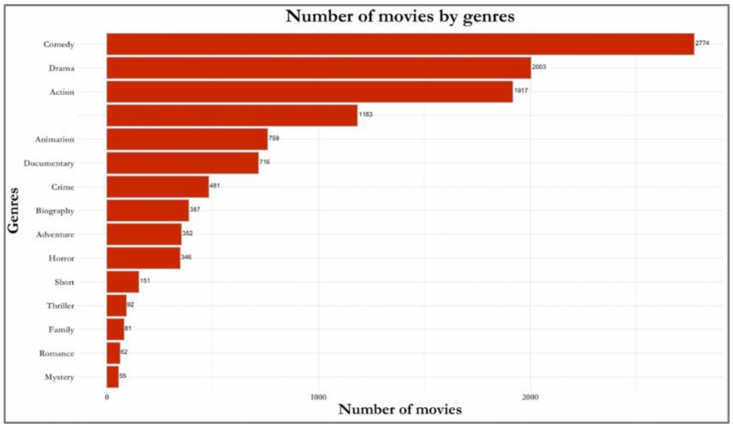
3.1. About the Data

The data was sourced from an online open community called Kaggle by Google. The link to the dataset is: www.kaggle.com/datasets/ashishgup/netflix-rotten-tomatoes-metacritic-imdb. The dataset is diverse and contains over 15,000 movies across 29 genres, available in 100+ languages and 36 countries. The number of directors exceeds 7,000 and the number of actors exceeds 20,000.

[1] "Where is Netflix available?," Netflix Help Center, accessed 24 March 2022

[2] Amanda Silberling, "Netflix had its lowest year of subscriber growth since 2015," TechCrunch, 21 January 2022

Figure 1: Number of movies by genres



3.2. Data Preparation

Before diving into analysis, the data needed to be prepared for descriptive and inferential analysis. The team carried out the following to prepare the data –

1. Data Exclusions
   - 1.1. The dataset contained entries of movies and series. However, to limit the analysis to movies, around 3,700 entries with type of “Series” were excluded from the analysis. (3,690 records)
   - 1.2. Around 300 duplicate or inconsistent entries were excluded from the dataset for analysis. (296 records)
2. Additional derived data fields
   - 2.1. Primary Genre: the first value from the concatenated Genre field was considered as the Primary Genre of the movie
   - 2.2. Primary Director: the first value from the concatenated Director field was considered as the Primary Director of the movie
   - 2.3. Primary Language: The first value from the concatenated Director language field was considered as the Primary Language the movie was released in
   - 2.4. Genre Count: The total number of Genres tagged to a movie has been used to derive Genre Count variable
   - 2.5. Language Count: The total number of Languages tagged to a movie has been used to derive Language Count variable
   - 2.6. Country Count: The total number of Country tagged to a movie has been used to derive Country Count variable

Below is the list of libraries we has used for data analysis in R Studio:

Data Manipulation:	tidyverse, dplyr
Visualization:	ggplot2, ggthemes, scales, ggrepel, ggpubr, forcats, MASS, RcolorBrewer, viridis, tvthemes
Inferential Statistics:	ANOVA (shapiro.test, Tukey), Chi-Square (Chisq.test), Regression(lm), MASS, Corrplot

4. Descriptive Analysis

4.1. Data Filtering Criteria

The data fields that the team primarily analyzed are IMDb Ratings, Rotten Tomato Scores, Box Office Collection, and Awards Received against Genres, Actors, Directors, and Languages. However, given the limitations with the data (see section 6.1. for summary), the team had to set criteria to filter out noisy data. The criteria are –

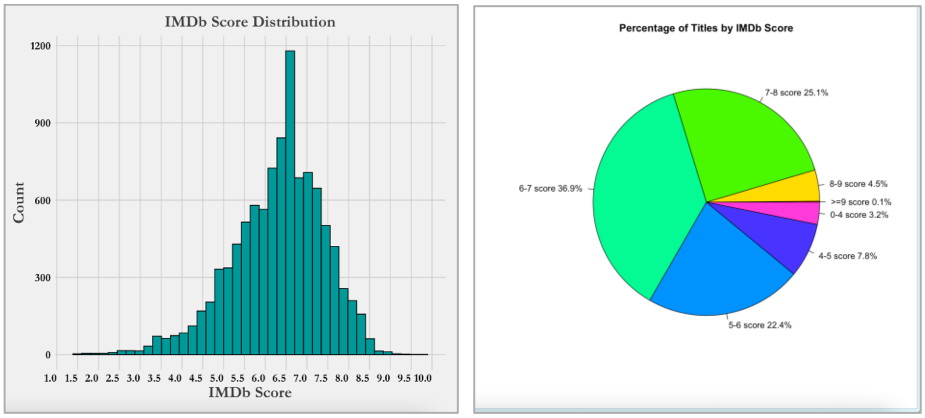
- Genres: The mean of IMDb Ratings, Rotten Tomato Scores and Box Office Collection, and cumulative of Awards received have been analyzed for genres with at least 20 movies in the dataset
- Actors: The mean of IMDb Ratings, Rotten Tomato Scores and Box Office Collection, and cumulative of Awards received have been analyzed for actors with at least 10 movies in the dataset
- Directors: The mean of IMDb Ratings, Rotten Tomato Scores and Box Office Collection, and cumulative of Awards received have been analyzed for directors with at least 10 movies in the dataset
- Languages: The mean of IMDb Ratings, Rotten Tomato Scores and Box Office Collection, and cumulative of Awards received have been analyzed for languages with at least 50 movies in the dataset

Additionally, for regression analysis, the dataset was further condensed as missing values significantly diluted the results. A set of around 2,500 movies that had full values. i.e., no missing values in any of the data fields and correctly recorded information were considered for regression.

4.2. IMDb Ratings

According to IMDb official website, IMDb score is aggregated and summarized from votes of individual registered users. In total, we obtained 10,075 movie titles that have IMDb scores. The most common IMDb score of movies on Netflix is from 6.5 to 7. This score range is not high given the score scale of 10. In fact, only 0.1% of the movies have scores of over 9 and 4.5% of movie scores are starting from 8. There are 10% of movies with ratings under the average of 5.

Figure 2. Distribution of IMDb Score and Proportion of Movies by IMDb ratings



Genre categorizes movies and can potentially evoke a wide spectrum of emotions on the audience. A mean analysis revealed that **Documentary**, **Biography**, **Short (movie), Crime** and **Animation** were the top five genres based on mean IMDb score. Famous documentaries such as No Festival & David Attenborough and Biographies such as Schindler's List can partly explain the high scores of these genres.

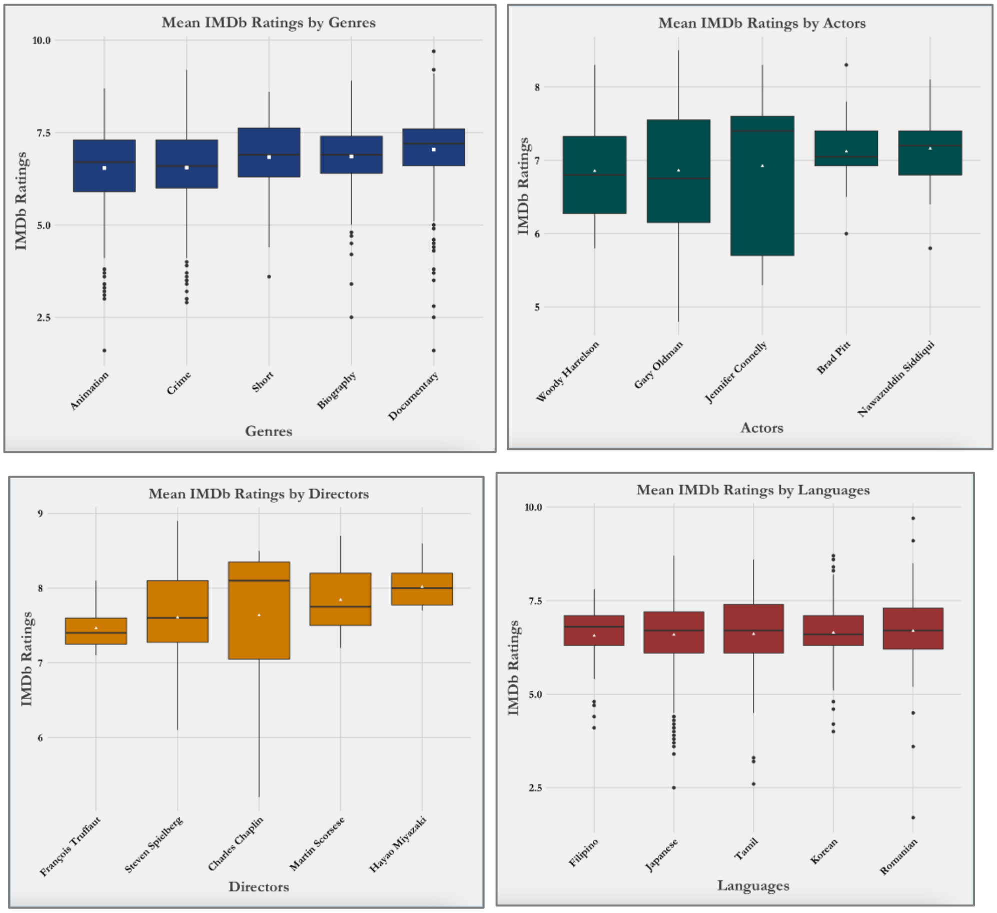
While comparing distributions of IMDb score across genres, we saw that IMDb distributions were left-skewed for many genres, including some in the top five such as **Documentary** and **Animation**. Given that IMDb ratings for movies from genres including Comedy, Action and **Drama** received low ratings, the overall IMDb ratings of these genres were significantly low compared to other genres. However, this pattern was not consistent across the **Mystery, Family and Fantasy** genre.

Actors and **Directors** play important roles in the success of movies. We wanted to explore the best actors and directors according to the ranking of IMDb users. We found the top five actors' list were mostly dominated by American actors, except **Nawazuddin Siddiqui**. The list also includes **Jennifer Connelly, Gary Oldman, Brad Pitt,** and **Woody Harrelson**. This weightage of ratings can be attributed to the breadth of viewership of Hollywood movies, globally. With respect to directors Hayao Miyazaki topped the top 5 with his animation movies. From the boxplot, we can observe his movies achieved a rating over 7.5. Thus, he is certainly amongst the best directors to guarantee a high score on IMDb. The other four directors in the top five are **Charles Chaplin, Martin Scorsese, Steven Spielberg,** and **François Truffaut.**

Overall, the distribution of IMDb ratings by actors and directors have just a few outliers compared to that by genre. Upper quartile outliers can be attributed to carefully made choices by actors and directors with respect to movies plots to maintain their high professional standards.

By **languages**, the top 5 languages that had the highest average IMDb score are Romanian, Korean, Tamil, Japanese, and Filipino.

Figure 3. Mean IMDb Ratings by Genres, Actors, Directors, and Languages

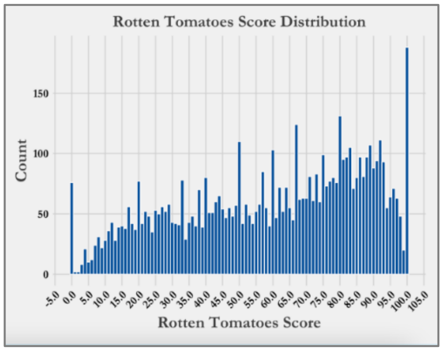


4.3. Rotten Tomato Scores

Rotten Tomatoes is one of the main competitors of IMDb as a review aggregator, hence it is important to assess the factors that differentiate these two entities. The 'tomato meter' has a score ranging from 0 to 100. Based on the distribution chart of Rotten Tomato, there is no discernable distribution followed by the scores. Nonetheless we observed some peaks at score value 0, and 100, and other sporadic peaks in the middle. The data revealed that around 75, 131, 124, 111 and 188 movies in the dataset have secured a score of 0, 80, 67, 92, and 100, respectively on the tomato meter.

Although the distribution of Rotten Tomato score may differ from IMDb ratings, their mean values are similar – 60 out of 100 for rotten tomatoes and 6.3 out of 10 for IMDb.

Figure 4. Distribution of Rotten Tomato Scores



The top five **Genres** in order are Documentary, Biography, Animation, Drama and Crime. Fifty-two movies including but not limited to Bill Nye: Science Guy, Waste Land and Infinite Football in the *Documentary* genre have secured 100 rating on the tomato meter, driving the mean score towards the upper range. Some of the top-rated movies in the *Crime* genre are A Bright Summer Day, Nigerian Prince, Vengeance is Mine. For *Drama*, it was Lola, Kabhi Khushi Kabhi Gham and Tokyo Story. For *Animation*, the top drivers were Fatal Five, Seoul Station, Penguin Highway, and DC's Justice League.

The top-rated **Actors** were French superstar *Jean-Paul Belmondo*, Devil Wears Prada's fame *Emily Blunt*, two Hindi movie (part of Indian movie industry) superstars – one, most critically acclaimed *Nawazuddin Siddiqui* and another, now global icon, *Priyanka Chopra*, and the veteran Chinese actor *Maggie Cheung*.

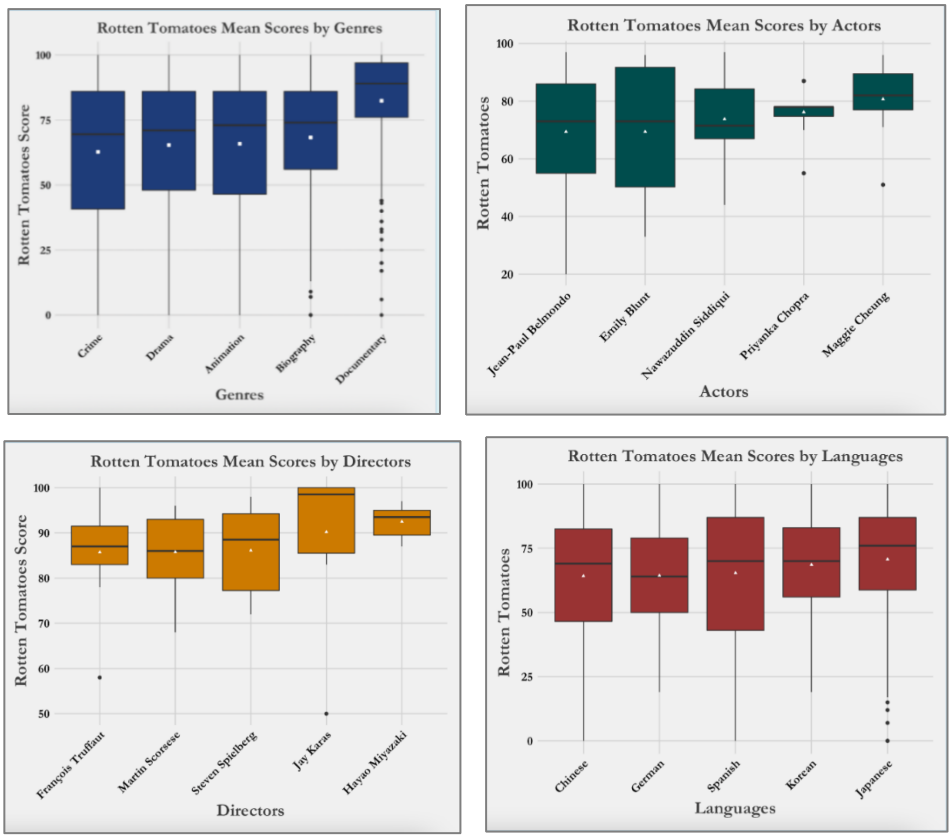
Although *Priyanka Chopra* has appeared in at least 67 movies[3] including Hindi and English, this dataset records only a limited section of Priyanka Chopra's career till date.

With respect to **Directors**, *Hayao Miyazaki, Jay Karas, Steven Spielberg, Martin Scorsese, and François Truffaut*. While *Karas* and *Spielberg* outdid *Miyazaki* with respect to number of movies directed, *Miyazaki*, on average, was more critically acclaimed vis-à-vis other directors in the list.

In terms of **Languages**, *Japanese, Korean, Spanish, German, Chinese*, topped the mean Rotten Tomato Score chart. While *Chinese and Spanish* movies outdid the number of *Japanese*, *Japanese* movies were relatively more critically acclaimed.

[3] "List of Priyanka Chopra performances," Wikipedia, accessed 24 March 2022

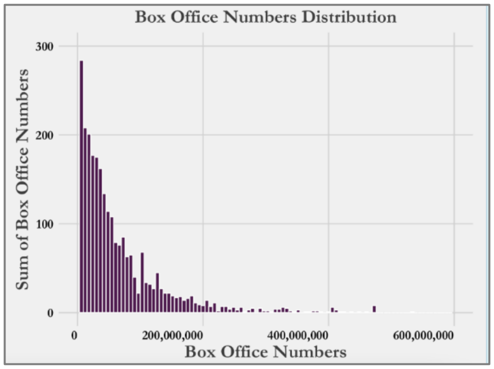
Figure 5. Mean Rotten Tomato Score by Genres, Actors, Directors, and Languages



4.4. Box Office

Box office values are an important metric to empirically assess the success of a movie. In our dataset, the distribution of box office values is severely right skewed. This is expected as only a handful number of movies "break" the box office collection.

Figure 6. Distribution of Box Office Numbers



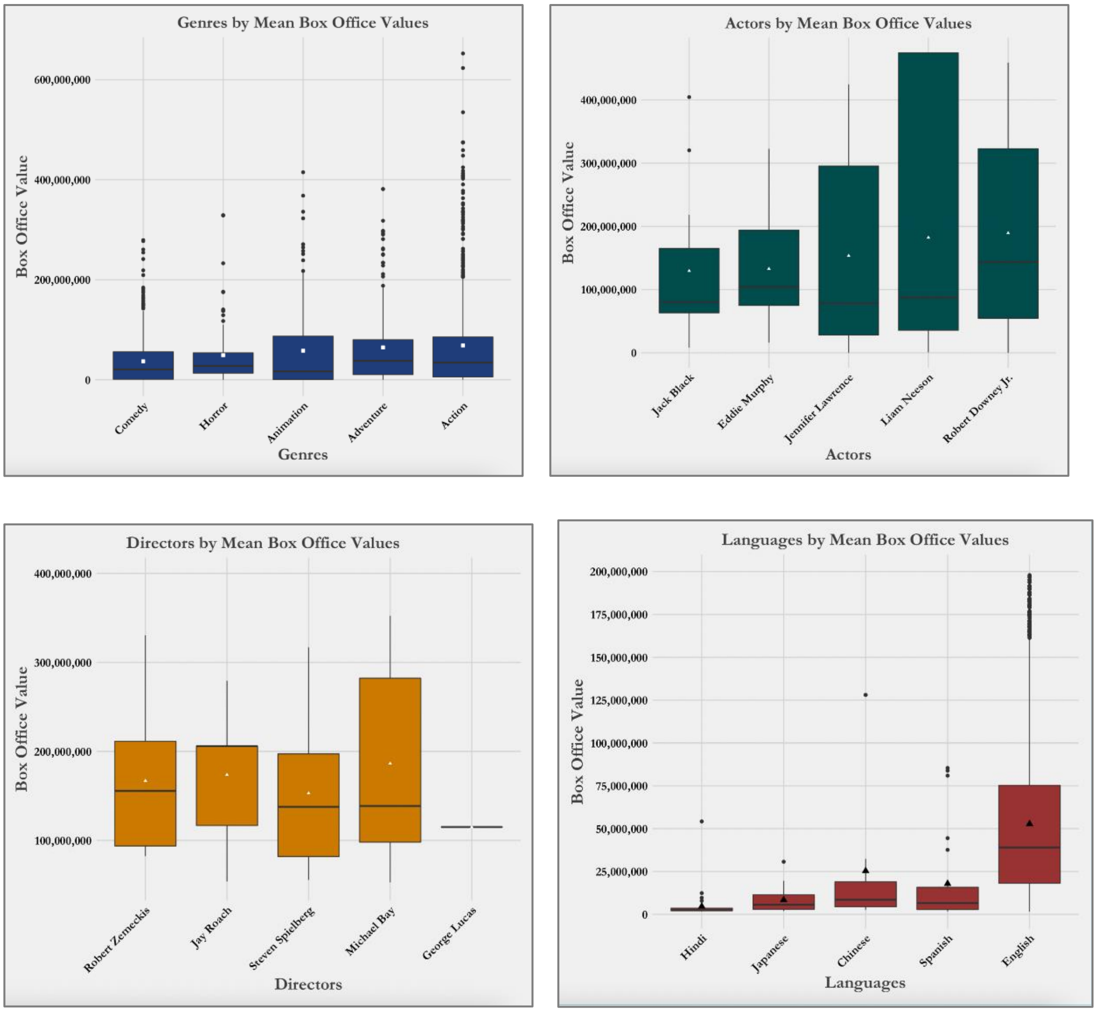
By **Genres,** *Action, Adventure, Animation, Horror and Comedy* genre topped the box office by mean value chart. Each of the genres had a high number of outliers as well, especially *Action*. Marvel and DC movies as well as Jurassic Park series have been drivers of box office values for *Action* genre.

With respect to **Actors**, *Robert Downey Jr.* topped the mean box office chart, even though the cumulative value of box office collection for *Liam Neeson* was greater than *Downey*. This is because *Neeson* has appeared in many films that have broken box office records more than a decade ago, which when adjusted by inflation probably could have been at par with *Downey.* In the last decade, *Downey's* popularity as Iron Man and Sherlock Holmes has resulted in grossing higher cumulative box office values even with a lower number of movies compared with *Neeson*.

With respect to **Directors**, George Lucas with a handful of movies, far exceeded the mean box office collection compared to other directors, including *Michael Bay, Steven Spielberg, Jay Roach, and Zemeckis*.

With respect to **Languages**, the difference in the currency value and viewership breadth are two biggest factors resulting in a significant difference in the mean box collection of *English* language movies vis-à-vis other top listed languages including *Spanish, Chinese, Japanese, and Hindi*.

Figure 7. Mean Box Office Values by Genres, Actors, Directors, and Languages



4.5. Awards Received

As per our analysis, 40% of the movies in our dataset received at least one award. Amongst those receiving at least one award, 25% received 10 awards or more.

Awards are a paramount metric indicating success for a movie, an actor, and a director. The higher the number of awards actors and directors have in their career and movies have in their heyday, the more successful they may be. For this reason, we deemed that it is appropriate to compare genres, directors, and actors by the cumulative of awards achieved by them rather than by the average value of awards they have received.

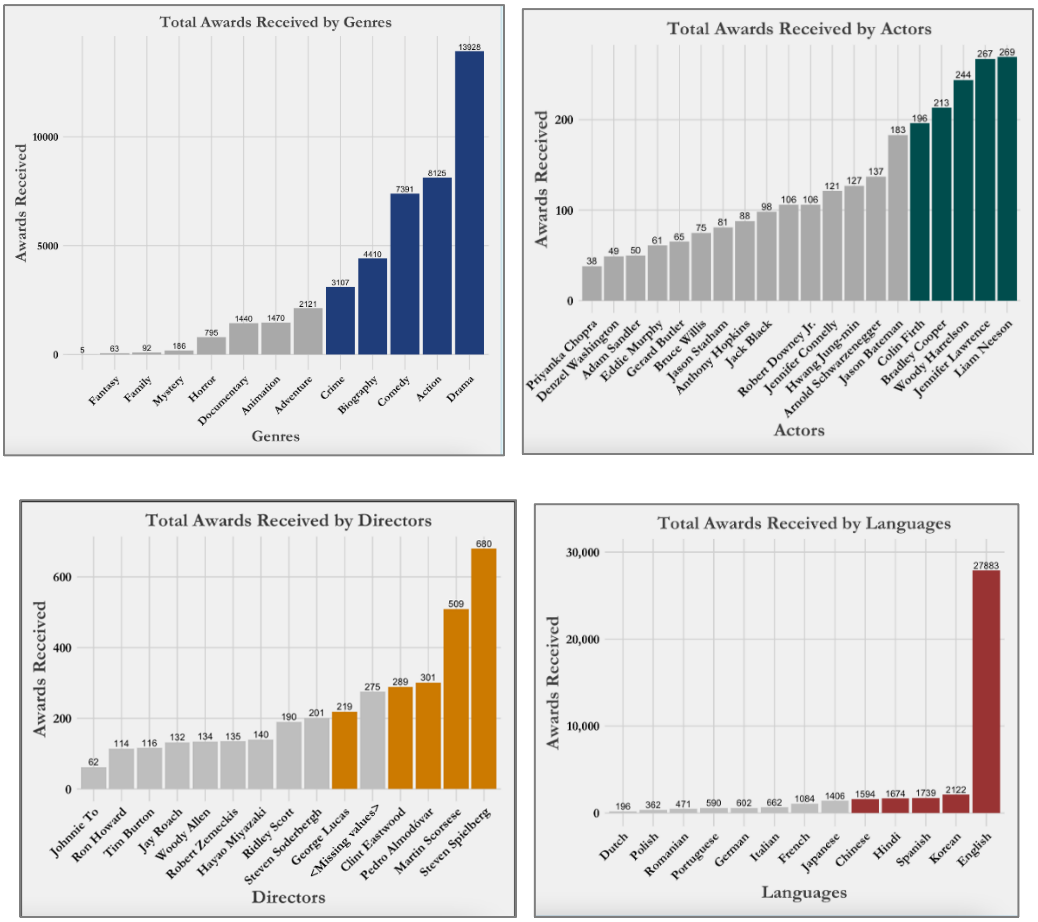
By **Genre**, we found that *Drama* did exceedingly well as compared to the other genres. *Action, Comedy, Biography, and Crime* were the other genres that topped the charts with respect to cumulative awards received. Such a high value can be potentially attributed to movies such as Parasite, Roma, and Gravity that were well received by viewers globally.

Actors *Liam Neeson, Jennifer Lawrence, Woody Harrelson, Bradley Cooper, and Collin Firth* grabbed the highest number of awards. *Liam Neeson's* tenure in the entertainment business and his role in well renowned series such as Taken, and Star Wars are potential factors to his success. For Jennifer Lawrence, the same can be credited to the popularity of Silver Linings Playbook and The Hunger Games globally.

With respect to **Directors**, *Steven Spielberg* topped the list of cumulative awards received. *Steven Spielberg* is a globally touted director in the movie industry. It is thus not surprising to see that *Spielberg* outdid the number of awards received, followed by *Martin Scorsese*. Global success of *Spielberg's* movies such as E.T., Indiana Jones series, and Schindler's List can be attributed to cumulative awards he has received. The other directors that featured in the list include *Pedro Almodovar, Clint Eastwood, and George Lucas.*

In **Languages**, the number of awards received by *English* movies far exceeded those received by other languages. One factor can be credited to the breadth of viewership of *English* movies, and the other can be attributed to the IMDb database being primarily dominated by *English* cinema awards.

Figure 8. Total number of awards received by Genres, Actors, Directors, and Languages



5. Inferential Analysis

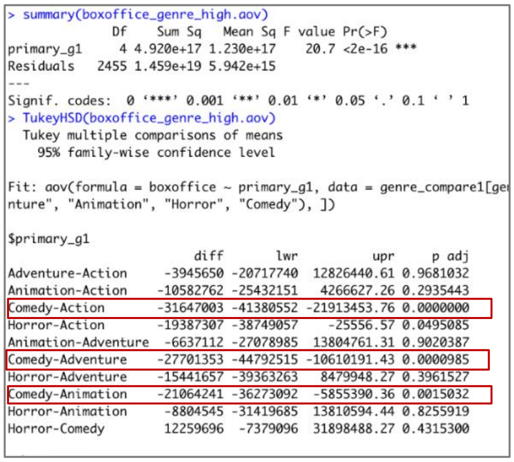
5.1. Analysis of Variance

ANOVA analysis is performed for Box Office, IMDb score, Rotten Tomatoes score against genres, actors, directors, and languages. In the report, we analyze only those with meaningful results. The insignificant results have been populated in Appendix (see section 7) for reference.

1. ANOVA for top five genres by Box Office Numbers

   H_0: There is no difference in the mean IMDb score of top five genres
   H_A: There is at least one genre out of five that has a different mean IMDb score

Figure 9. ANOVA to test difference in the IMDb mean score of top five genres



```
> summary(boxoffice_genre_high.aov)
              Df    Sum Sq   Mean Sq F value Pr(>F)
primary_g1     4 4.920e+17 1.230e+17    20.7 <2e-16 ***
Residuals   2455 1.459e+19 5.942e+15
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(boxoffice_genre_high.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = boxoffice ~ primary_g1, data = genre_compare1[ge
nture", "Animation", "Horror", "Comedy"), ])

$primary_g1
                         diff       lwr          upr     p adj
Adventure-Action     -3945650 -20717740  12826440.61 0.9681032
Animation-Action    -10582762 -25432151   4266627.26 0.2935443
Comedy-Action       -31647003 -41380552 -21913453.76 0.0000000
Horror-Action       -19387307 -38749057    -25556.57 0.0495085
Animation-Adventure  -6637112 -27078985  13804761.31 0.9020387
Comedy-Adventure    -27701353 -44792515 -10610191.43 0.0000985
Horror-Adventure    -15441657 -39363263   8479948.27 0.3961527
Comedy-Animation    -21064241 -36273092  -5855390.36 0.0015032
Horror-Animation     -8804545 -31419685  13810594.44 0.8255919
Horror-Comedy        12259696  -7379096  31898488.27 0.4315300
```

Since the p value < 0.05, we reject the null hypothesis and conclude that there is a difference in IMDb score across top highest-scoring genres.

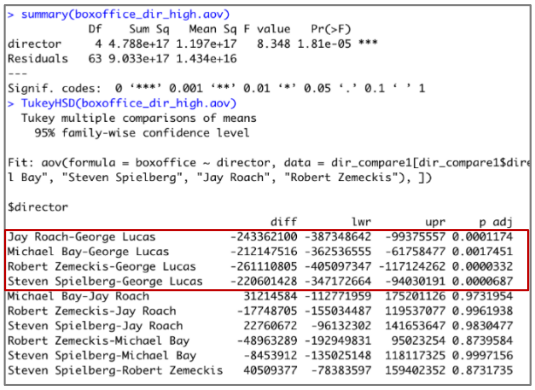
Tukey's HSD helps indicate the difference in the mean of -

- Comedy – Action
- Comedy – Animation
- Comedy – Adventure

2. ANOVA for top five directors by Box Office Numbers

   H_0: There is no difference in the mean box office numbers of top five directors
   H_A: There is at least one director out of five that has a different mean box office value

Figure 10. ANOVA to test difference in the mean box office values of top five directors



```
> summary(boxoffice_dir_high.aov)
            Df    Sum Sq   Mean Sq F value   Pr(>F)
director     4 4.788e+17 1.197e+17   8.348 1.81e-05 ***
Residuals   63 9.033e+17 1.434e+16
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(boxoffice_dir_high.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = boxoffice ~ director, data = dir_compare1[dir_compare1$dire
l Bay", "Steven Spielberg", "Jay Roach", "Robert Zemeckis"), ])

$director
                                  diff        lwr        upr     p adj
Jay Roach-George Lucas          -243362100 -387348642  -99375557 0.0001174
Michael Bay-George Lucas        -212147516 -362536555  -61758477 0.0017451
Robert Zemeckis-George Lucas    -261110805 -405097347 -117124262 0.0000332
Steven Spielberg-George Lucas   -220601428 -347172664  -94030191 0.0000687
Michael Bay-Jay Roach             31214584 -112771959  175201126 0.9731954
Robert Zemeckis-Jay Roach        -17748705 -155034487  119537077 0.9961938
Steven Spielberg-Jay Roach        22760672  -96132302  141653647 0.9830477
Robert Zemeckis-Michael Bay      -48963289 -192949831   95023254 0.8739584
Steven Spielberg-Michael Bay      -8453912 -135025148  118117325 0.9997156
Steven Spielberg-Robert Zemeckis  40509377  -78383597  159402352 0.8731735
```

Since the p value < 0.05, we reject the null hypothesis and conclude that there is a difference in mean box office value of top 5 directors.

Tukey's HSD helps indicate the difference in the mean of -

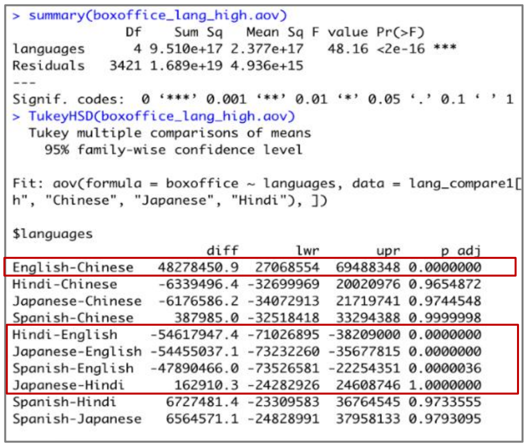
- Jay Roach – George Lucas
- Micheal Bay - George Lucas
- Robert Zemeckis – George Lucas
- Steven Spielberg – George Lucas

Clearly, George Lucas' mean box office value outdid the other directors in the list.

3. ANOVA for top five languages by Box Office Numbers

H_0: There is no difference in the mean box office values of the top five languages
H_A: There is at least one language out of five that has a different mean box office value

Figure 11. ANOVA to test difference in the mean box office value of top five languages



```
> summary(boxoffice_lang_high.aov)
              Df    Sum Sq   Mean Sq F value Pr(>F)    
languages      4 9.510e+17 2.377e+17   48.16 <2e-16 ***
Residuals   3421 1.689e+19 4.936e+15                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(boxoffice_lang_high.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = boxoffice ~ languages, data = lang_compare1[
h", "Chinese", "Japanese", "Hindi"), ])

$languages
                        diff       lwr       upr     p adj
English-Chinese   48278450.9  27068554  69488348 0.0000000
Hindi-Chinese     -6339496.4 -32699969  20020976 0.9654872
Japanese-Chinese  -6176586.2 -34072913  21719741 0.9744548
Spanish-Chinese     387985.0 -32518418  33294388 0.9999998
Hindi-English    -54617947.4 -71026895 -38209000 0.0000000
Japanese-English -54455037.1 -73232260 -35677815 0.0000000
Spanish-English  -47890466.0 -73526581 -22254351 0.0000036
Japanese-Hindi      162910.3 -24282926  24608746 1.0000000
Spanish-Hindi      6727481.4 -23309583  36764545 0.9733555
Spanish-Japanese   6564571.1 -24828991  37958133 0.9793095
```

Since the p value < 0.05, we reject the null hypothesis and conclude that there is a difference in box office numbers across the top 5 languages.

Tukey's HSD helps indicate the difference in the mean of -

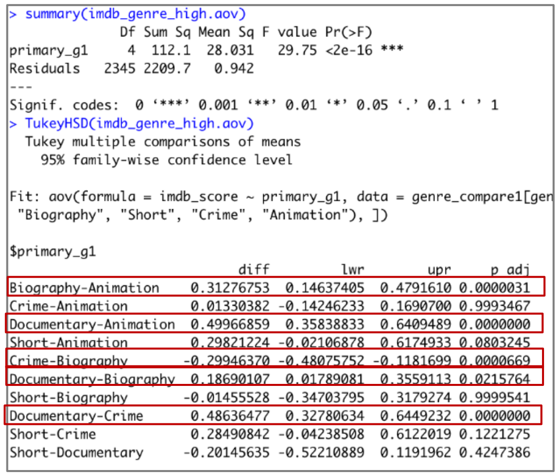
- English – Chinese
- Hindi – English
- Japanese – English
- Spanish – English
- Japanese – Hindi

Clearly, English language's mean box office value outdid those of other languages in the top list.

4. ANOVA for top five genres by IMDb scores

H_0: There is no difference in the mean IMDb scores of the top five genres
H_A: There is at least one genre out of five that has a different IMDb score

Figure 12. ANOVA to test the difference in mean IMDb ratings of top five genres



```
> summary(imdb_genre_high.aov)
              Df Sum Sq Mean Sq F value Pr(>F)    
primary_g1     4  112.1  28.031   29.75 <2e-16 ***
Residuals   2345 2209.7   0.942                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(imdb_genre_high.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = imdb_score ~ primary_g1, data = genre_compare1[gen
 "Biography", "Short", "Crime", "Animation"), ])

$primary_g1
                            diff         lwr        upr     p adj
Biography-Animation    0.31276753  0.14637405  0.4791610 0.0000031
Crime-Animation        0.01330382 -0.14246233  0.1690700 0.9993467
Documentary-Animation  0.49966859  0.35838833  0.6409489 0.0000000
Short-Animation        0.29821224 -0.02106878  0.6174933 0.0803245
Crime-Biography       -0.29946370 -0.48075752 -0.1181699 0.0000669
Documentary-Biography  0.18690107  0.01789081  0.3559113 0.0215764
Short-Biography       -0.01455528 -0.34703795  0.3179274 0.9999541
Documentary-Crime      0.48636477  0.32780634  0.6449232 0.0000000
Short-Crime            0.28490842 -0.04238508  0.6122019 0.1221275
Short-Documentary     -0.20145635 -0.52210889  0.1191962 0.4247386
```

Since the p value < 0.05, we reject the null hypothesis and conclude that there is a difference in IMDb scores across the top five genres.

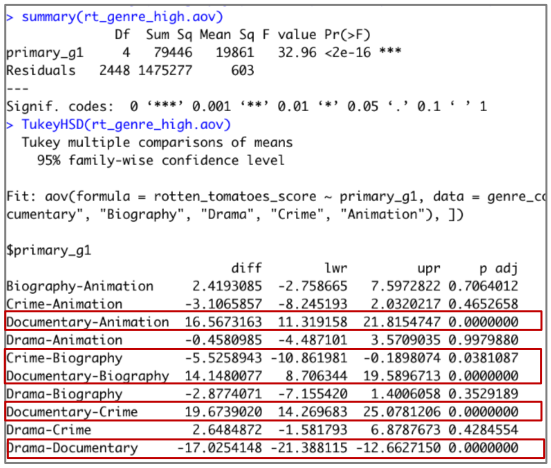
Tukey’s HSD helps indicate the difference in the mean of -

- Biography – Animation
- Documentary – Animation
- Crime – Biography
- Documentary – Biography
- Documentary – Crime

5. ANOVA for top five genres by Rotten Tomatoes score

H$_0$: There is no difference in the mean Rotten Tomatoes score of the top five genres
H$_A$: There is at least one genre out of five that has a different Rotten Tomatoes Score

Figure 13. ANOVA to test difference in the mean Rotten Tomatoes score of top five genres



```
> summary(rt_genre_high.aov)
              Df  Sum Sq Mean Sq F value Pr(>F)    
primary_g1     4   79446   19861   32.96 <2e-16 ***
Residuals   2448 1475277     603                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(rt_genre_high.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = rotten_tomatoes_score ~ primary_g1, data = genre_co
cumentary", "Biography", "Drama", "Crime", "Animation"), ])

$primary_g1
                            diff        lwr         upr     p adj
Biography-Animation     2.4193085  -2.758665   7.5972822 0.7064012
Crime-Animation        -3.1065857  -8.245193   2.0320217 0.4652658
Documentary-Animation  16.5673163  11.319158  21.8154747 0.0000000
Drama-Animation        -0.4580985  -4.487101   3.5709035 0.9979880
Crime-Biography        -5.5258943 -10.861981  -0.1898074 0.0381087
Documentary-Biography  14.1480077   8.706344  19.5896713 0.0000000
Drama-Biography        -2.8774071  -7.155420   1.4006058 0.3529189
Documentary-Crime      19.6739020  14.269683  25.0781206 0.0000000
Drama-Crime             2.6484872  -1.581793   6.8787673 0.4284554
Drama-Documentary     -17.0254148 -21.388115 -12.6627150 0.0000000
```

Since the p value < 0.05, we reject the null hypothesis and conclude that there is a difference in the rotten tomatoes score across the top five genres.

Tukey’s HSD helps indicate the difference in the mean of -

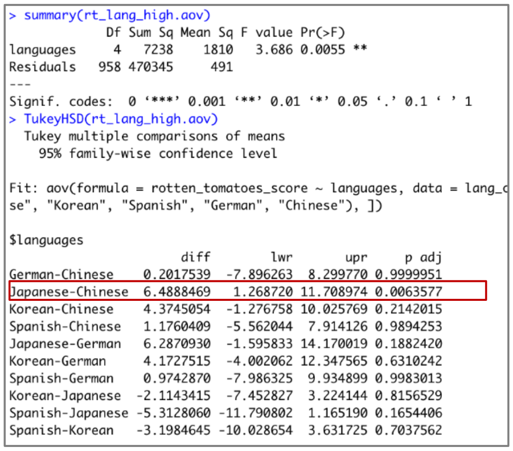
- Documentary – Biography

- Documentary – Animation
- Crime – Biography
- Documentary – Crime
- Drama – Documentary

6. ANOVA for top five languages by Rotten Tomatoes score

   H_0: There is no difference in the mean Rotten Tomatoes score of the top five languages
   H_A: There is at least one language out of five that has a different Rotten Tomatoes Score

Figure 14. ANOVA to test the difference in mean Rotten Tomatoes score of top five languages



```
> summary(rt_lang_high.aov)
             Df Sum Sq Mean Sq F value Pr(>F)   
languages     4   7238    1810   3.686 0.0055 **
Residuals   958 470345     491                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(rt_lang_high.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = rotten_tomatoes_score ~ languages, data = lang_d
se", "Korean", "Spanish", "German", "Chinese"), ])

$languages
                       diff        lwr       upr     p adj
German-Chinese    0.2017539  -7.896263  8.299770 0.9999951
Japanese-Chinese  6.4888469   1.268720 11.708974 0.0063577
Korean-Chinese    4.3745054  -1.276758 10.025769 0.2142015
Spanish-Chinese   1.1760409  -5.562044  7.914126 0.9894253
Japanese-German   6.2870930  -1.595833 14.170019 0.1882420
Korean-German     4.1727515  -4.002062 12.347565 0.6310242
Spanish-German    0.9742870  -7.986325  9.934899 0.9983013
Korean-Japanese  -2.1143415  -7.452827  3.224144 0.8156529
Spanish-Japanese -5.3128060 -11.790802  1.165190 0.1654406
Spanish-Korean   -3.1984645 -10.028654  3.631725 0.7037562
```

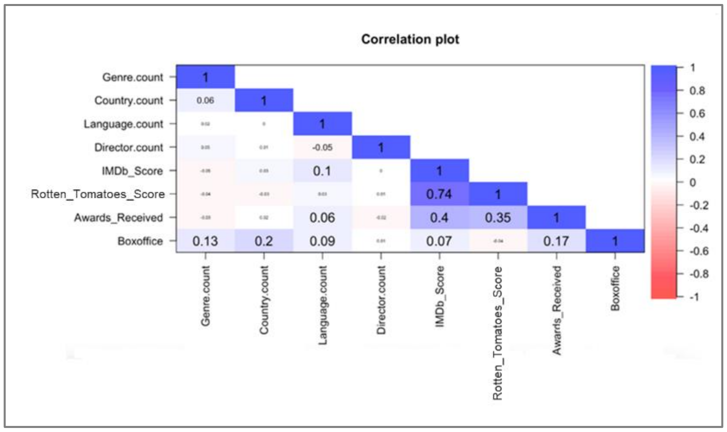
Since the p value < 0.05, we reject the null hypothesis and conclude that there is a difference in the rotten tomatoes score across the top 5 languages.

Tukey’s HSD helps indicate the difference in the mean of -

- Japanese – Chinese

5.2. Correlation and Regression Analysis

Figure 15: Correlation analysis of continuous variables



The correlation between IMDb ratings and Rotten Tomatoes score is high at 0.74 whereas the correlation between Rotten Tomatoes scores and Awards Received is weak.

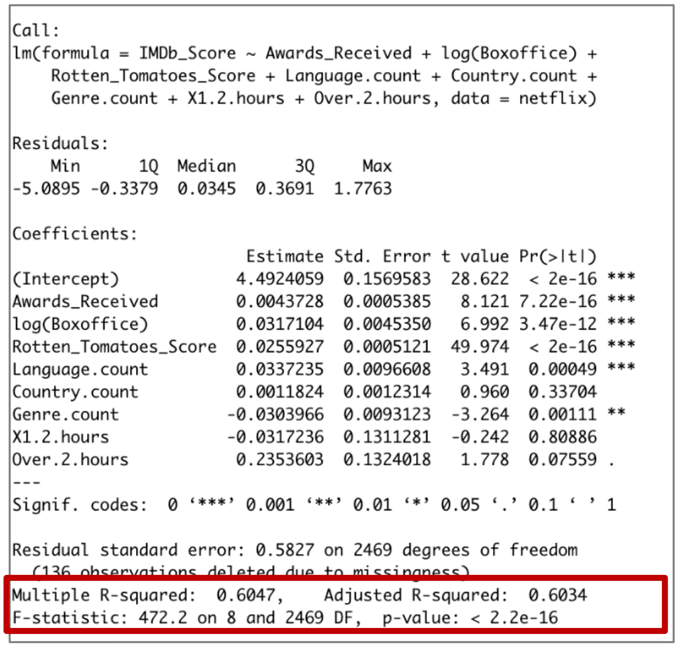
We ran a regression analysis with Box Office value, Rotten Tomatoes Scores and IMDb Ratings as the dependent variables. We also ran Stepwise regression with the stopping rule as minimum AIC decrease the complexity of the model and reduce the number of predictors without compromising on the model performance. However, the difference in the results from the linear regression and the stepwise regression is negligible and thus we retained the output from the former analysis.

1. **IMDb Ratings prediction**

   H_0: There is no significant linear relationship between IMDb score and the independent variables
   H_A: There is a significant linear relationship between the IMDb score and the independent variables

   lm(IMDb_Score ~ Awards_Received + log(Boxoffice) + Rotten_Tomatoes_Score + Language.count + Country.count + Genre.count + X1.2.hours + Over.2.hours, data = Netflix_dataset)

Figure 16. Linear regression analysis for IMDb Ratings



```
Call:
lm(formula = IMDb_Score ~ Awards_Received + log(Boxoffice) + 
    Rotten_Tomatoes_Score + Language.count + Country.count + 
    Genre.count + X1.2.hours + Over.2.hours, data = netflix)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.0895 -0.3379  0.0345  0.3691  1.7763 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            4.4924059  0.1569583  28.622  < 2e-16 ***
Awards_Received        0.0043728  0.0005385   8.121 7.22e-16 ***
log(Boxoffice)         0.0317104  0.0045350   6.992 3.47e-12 ***
Rotten_Tomatoes_Score  0.0255927  0.0005121  49.974  < 2e-16 ***
Language.count         0.0337235  0.0096608   3.491  0.00049 ***
Country.count          0.0011824  0.0012314   0.960  0.33704    
Genre.count           -0.0303966  0.0093123  -3.264  0.00111 ** 
X1.2.hours            -0.0317236  0.1311281  -0.242  0.80886    
Over.2.hours           0.2353603  0.1324018   1.778  0.07559 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5827 on 2469 degrees of freedom
  (136 observations deleted due to missingness)
Multiple R-squared:  0.6047,	Adjusted R-squared:  0.6034 
F-statistic: 472.2 on 8 and 2469 DF,  p-value: < 2.2e-16
```

Since the p value < 0.05, we reject the null hypothesis and conclude that there is a significant linear relationship between the independent variables and the IMDb Ratings.

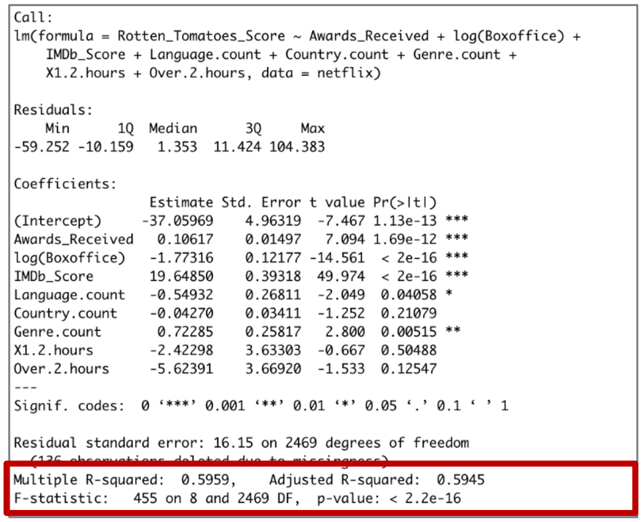
Result: Awards Received, Box Office Value, Rotten Tomatoes Score, Language Count, and Genre count have a significant influence on the IMDb score of the movie. The R square value is close to 60%, which means that 60% variance in the IMDb score can be explained by the independent variables.

2. **Rotten Tomatoes Score prediction**

   H_0: There is no significant relationship between Rotten Tomatoes score and the independent
   H_A: There is a significant linear relationship between the Rotten Tomatoes Scores and the independent variables

   lm(Rotten_Tomatoes_Score ~ Awards_Received + log(Boxoffice) + IMDb_Score + Language.count + Country.count + Genre.count + X1.2.hours + Over.2.hours, data = Netflix_dataset)

Figure 17. Linear regression analysis for Rotten Tomatoes Scores



```
Call:
lm(formula = Rotten_Tomatoes_Score ~ Awards_Received + log(Boxoffice) +
    IMDb_Score + Language.count + Country.count + Genre.count +
    X1.2.hours + Over.2.hours, data = netflix)

Residuals:
    Min      1Q  Median      3Q     Max
-59.252 -10.159   1.353  11.424 104.383

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -37.05969    4.96319  -7.467 1.13e-13 ***
Awards_Received   0.10617    0.01497   7.094 1.69e-12 ***
log(Boxoffice)   -1.77316    0.12177 -14.561  < 2e-16 ***
IMDb_Score       19.64850    0.39318  49.974  < 2e-16 ***
Language.count   -0.54932    0.26811  -2.049  0.04058 *
Country.count    -0.04270    0.03411  -1.252  0.21079
Genre.count       0.72285    0.25817   2.800  0.00515 **
X1.2.hours       -2.42298    3.63303  -0.667  0.50488
Over.2.hours     -5.62391    3.66920  -1.533  0.12547
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.15 on 2469 degrees of freedom
  (136 observations deleted due to missingness)
Multiple R-squared:  0.5959,    Adjusted R-squared:  0.5945
F-statistic:   455 on 8 and 2469 DF,  p-value: < 2.2e-16
```

Since the p value < 0.05, we reject the null hypothesis and conclude that there is a significant linear relationship between the and Rotten Tomato Scores.

Result: Awards Received, Box Office Values, IMDb Ratings, Language Count, and Genre Count has a significant influence on the Rotten Tomatoes score of the movie. The R square value is close to 60%, which means 60% of variation in the Rotten Tomatoes Score can be explained by the independent variables.

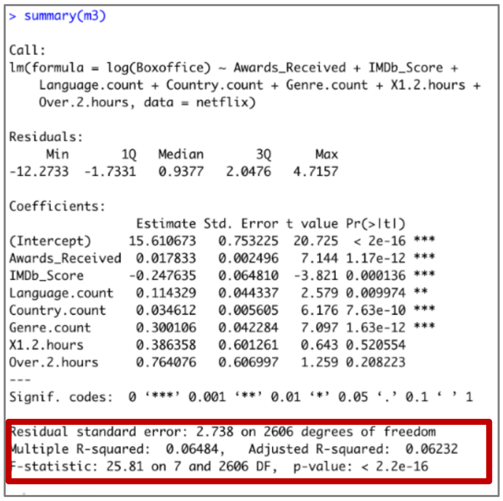
3. Box Office Value prediction

H_0: There is no significant linear relationship between Box Office Values and the independent variables
H_A: There is a significant linear relationship between the Box Office Values and the independent variables

lm(log(Boxoffice) ~ Awards_Received + IMDb_Score + Language.count + Country.count + Genre.count + X1.2.hours + Over.2.hours, data = netflix)

Here, Rotten Tomatoes score was removed from the model to avoid the issue of multicollinearity.

Figure 18. Linear regression analysis for Box Office Values



```
> summary(m3)

Call:
lm(formula = log(Boxoffice) ~ Awards_Received + IMDb_Score +
    Language.count + Country.count + Genre.count + X1.2.hours +
    Over.2.hours, data = netflix)

Residuals:
     Min       1Q   Median       3Q      Max
-12.2733  -1.7331   0.9377   2.0476   4.7157

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     15.610673   0.753225  20.725  < 2e-16 ***
Awards_Received  0.017833   0.002496   7.144 1.17e-12 ***
IMDb_Score      -0.247635   0.064810  -3.821 0.000136 ***
Language.count   0.114329   0.044337   2.579 0.009974 **
Country.count    0.034612   0.005605   6.176 7.63e-10 ***
Genre.count      0.300106   0.042284   7.097 1.63e-12 ***
X1.2.hours       0.386358   0.601261   0.643 0.520554
Over.2.hours     0.764076   0.606997   1.259 0.208223
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.738 on 2606 degrees of freedom
Multiple R-squared:  0.06484,   Adjusted R-squared:  0.06232
F-statistic: 25.81 on 7 and 2606 DF,  p-value: < 2.2e-16
```

Since the p value < 0.05, we reject the null hypothesis and conclude that there are some independent variables that contribute and should remain in the model.

Result: Awards received, IMDb score, language count and genre count have a significant influence on the box office numbers of the movie. However, **since the R square value is too low i.e., 6%, we can say that the independent variables cannot statistically explain significant variation in the Box Office Values.**

5.3. Chi Square Analysis

In the previous section, we explored prediction models for IMDb Ratings, Rotten Tomatoes Scores and Box Office Values. Additionally, the number of awards received is a reliable measure for the quality of movies.

We wanted to know if we can predict the IMDb Ratings or Box Office Value of a movie, and whether that information can tell us anything about the Number of Awards that a movie may get. With the belief that user ratings and box office possibly reflect the evaluation of movie critics, we conducted the association test between IMDb Score and Total Number of Awards as well as Box Office Collection and Total Number of awards.

H_0: There is no association between IMDb score and Number of Awards
H_A: There is an association between IMDb score and Number of Awards

Since X^2 = 1137.7, and the p-value < 0.05, we reject the null hypothesis. Thus, we have enough statistical evidence to conclude that there is an association between IMDb Score and Number of Awards Received.

H_0: There is no association between Box Office and Number of Awards
H_A: There is an association between Box Office and Number of Awards

Since X^2 = 130.62 and the p-value < 0.05, we reject the null hypothesis. Thus, we have enough statistical evidence to conclude that there is an association between Box Office and Number of Awards Received.

6. Conclusion

6.1. Limitations

Given that the dataset is sourced from a public domain Kaggle, the data has some limitations. These include -

1. Missing data – Out of over 11,000 movie entries, only around 6,000 movies had Rotten Tomato scores and around 3,800 movies had box office values.

2. Multiselect data – Multiselect data field such as genres – that had a range of 1 to 12 genres tags for a movie – made running analysis complex and limited.

3. Limited data fields – Data fields containing information that whether a movie aired in theatres or exclusively on Netflix platform would have resulted in team conducting hypothesis testing.

4. Reliability of data – As the data is sourced and uploaded by individual contributors there may be some errors or discrepancies to the true values. Also, the content on Netflix varies across region on availability dependent on licenses to stream on the service.

5. Skewness in consumer preference – In many regions the consumer may not depend on IMDb score or any metric to choose a movie, they may simply be fans of certain actors or have preference to a genre that directly influences their decision making.

6.2. Recommendations:

From the various findings detailed in this report, the consumer can make a much more informed decision when choosing a film. For Netflix, and movie producers, directors and actors, the statistics guide their decision-making process when considering making a movie and the factors that would contribute to the high ratings which usually translates to high box office collection, number of awards received, and in turn, greater recognition.

First, from the consumers' perspective and desire to watch a good movie, there are a few factors that can contribute to the decision-making process. As shown in Figure 3. IMDb Ratings by Genre, Actors, Directors, and Languages, critically acclaimed films tend to be from Drama genre. And if the viewer prefers to select a movie based on the Actor, it is best to choose a movie which contains Liam Neeson (male), Robert Downey Jr. (male), or Jennifer Lawrence (female) in its cast.

If the viewer would prefer to select a movie based on reviews on acclaimed sites such as Rotten Tomatoes, the data suggests choosing from the Documentary genre and in Japanese language as these tend to be highly rated on the tomato meter.

Next, for movie producers the findings are relevant and can help guide them depending on whether the target is to produce an award-clencher or a box-office driver. To achieve better success in terms of awards, it is advisable to produce an English movie, with Steven Spielberg or Martin Scorsese as the director. The genre could be either Drama or Action.

From the analysis we can see that there are clear patterns depending on the outcome we aim to achieve.

6.3. Ideas for further study

This study covers a large database of movies on Netflix, however there is room for improvement.

With much more data available in viewer preferences and viewing habits, having behavioral data will help with understanding why certain movies tend to score low in box-office earnings but perform better on streaming platforms. This could be useful for producers who may simply want to make movies for streaming platforms only.

Across different regions, the cultural and language preferences may differ and having region segmented data would provide valuable insights. Since 2021, Netflix reports top 10 English and non-English shows on a weekly basis.[4] This further incentivizes producers and intrigues viewers to search for the right titles to spend time and money on.

Outside of viewer preferences and movie data, combining multiple datasets such as world news and trends pertaining to the entertainment industry may indicate greater correlations that can aid in enhancing movies-related decision-making and success predictions.

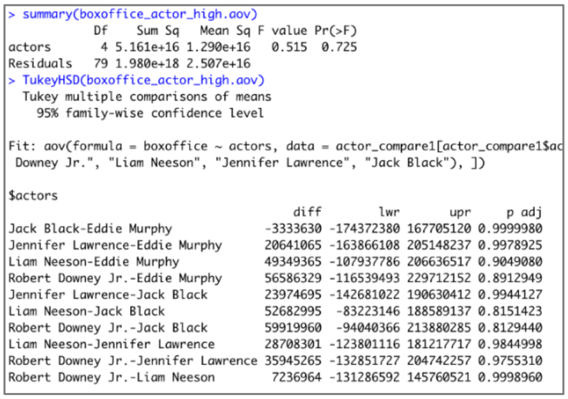
[4] "Netflix to release top 10 reports weekly for English and non-English TV and films," NBC News, 17 November 2021

6.4. Closing Statement

The study of movie data on Netflix has provided fascinating insights into an industry that moves with and shapes conversations in our lives. While we have sufficient findings to be able to recommend movie titles to viewers, and a direction to producers with the applied statistical methods, there are also many more methods and datasets that could potentially augment our findings and provide enhanced details with greater certainty. This study whilst substantial at this point has further room for imagination and growth in the future.

7. Appendix

ANOVA analysis that do not yield significant results have included below for perusal.

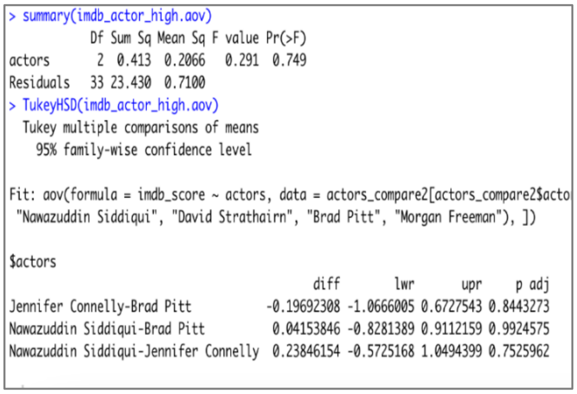


```
> summary(boxoffice_actor_high.aov)
            Df    Sum Sq   Mean Sq F value Pr(>F)
actors       4 5.161e+16 1.290e+16   0.515  0.725
Residuals   79 1.980e+18 2.507e+16
> TukeyHSD(boxoffice_actor_high.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = boxoffice ~ actors, data = actor_compare1[actor_compare1$ac
 Downey Jr.", "Liam Neeson", "Jennifer Lawrence", "Jack Black"), ])

$actors
                                     diff        lwr       upr     p adj
Jack Black-Eddie Murphy          -3333630 -174372380 167705120 0.9999980
Jennifer Lawrence-Eddie Murphy   20641065 -163866108 205148237 0.9978925
Liam Neeson-Eddie Murphy         49349365 -107937786 206636517 0.9049080
Robert Downey Jr.-Eddie Murphy   56586329 -116539493 229712152 0.8912949
Jennifer Lawrence-Jack Black     23974695 -142681022 190630412 0.9944127
Liam Neeson-Jack Black           52682995  -83223146 188589137 0.8151423
Robert Downey Jr.-Jack Black     59919960  -94040366 213880285 0.8129440
Liam Neeson-Jennifer Lawrence    28708301 -123801116 181217717 0.9844998
Robert Downey Jr.-Jennifer Lawrence 35945265 -132851727 204742257 0.9755310
Robert Downey Jr.-Liam Neeson     7236964 -131286592 145760521 0.9998960
```

Figure: Box office vs Actors

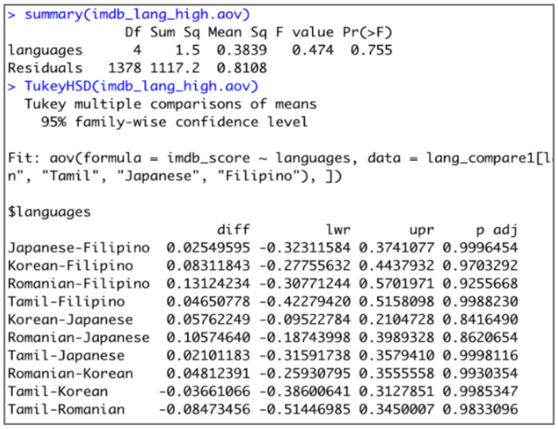


```
> summary(imdb_actor_high.aov)
            Df Sum Sq Mean Sq F value Pr(>F)
actors       2  0.413  0.2066   0.291  0.749
Residuals   33 23.430  0.7100
> TukeyHSD(imdb_actor_high.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = imdb_score ~ actors, data = actors_compare2[actors_compare2$acto
 "Nawazuddin Siddiqui", "David Strathairn", "Brad Pitt", "Morgan Freeman"), ])

$actors
                                          diff        lwr       upr     p adj
Jennifer Connelly-Brad Pitt         -0.19692308 -1.0666005 0.6727543 0.8443273
Nawazuddin Siddiqui-Brad Pitt        0.04153846 -0.8281389 0.9112159 0.9924575
Nawazuddin Siddiqui-Jennifer Connelly 0.23846154 -0.5725168 1.0494399 0.7525962
```

Figure: IMDb Score vs Actors

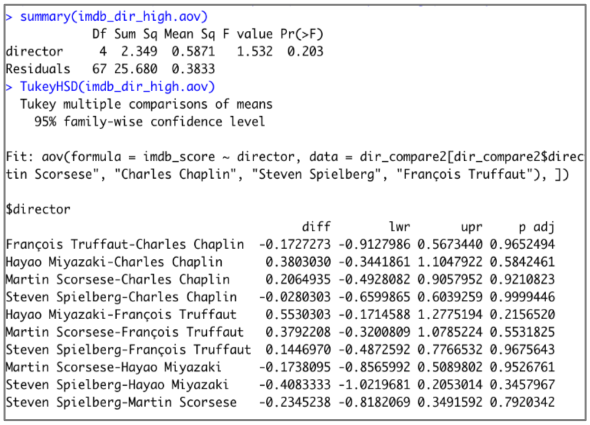


```
> summary(imdb_lang_high.aov)
              Df Sum Sq Mean Sq F value Pr(>F)
languages      4    1.5  0.3839   0.474  0.755
Residuals   1378 1117.2  0.8108
> TukeyHSD(imdb_lang_high.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = imdb_score ~ languages, data = lang_compare1[l
n", "Tamil", "Japanese", "Filipino"), ])

$languages
                        diff         lwr       upr     p adj
Japanese-Filipino  0.02549595 -0.32311584 0.3741077 0.9996454
Korean-Filipino    0.08311843 -0.27755632 0.4437932 0.9703292
Romanian-Filipino  0.13124234 -0.30771244 0.5701971 0.9255668
Tamil-Filipino     0.04650778 -0.42279420 0.5158098 0.9988230
Korean-Japanese    0.05762249 -0.09522784 0.2104728 0.8416490
Romanian-Japanese  0.10574640 -0.18743998 0.3989328 0.8620654
Tamil-Japanese     0.02101183 -0.31591738 0.3579410 0.9998116
Romanian-Korean    0.04812391 -0.25930795 0.3555558 0.9930354
Tamil-Korean      -0.03661066 -0.38600641 0.3127851 0.9985347
Tamil-Romanian    -0.08473456 -0.51446985 0.3450007 0.9833096
```

Figure: IMDb Score vs Language

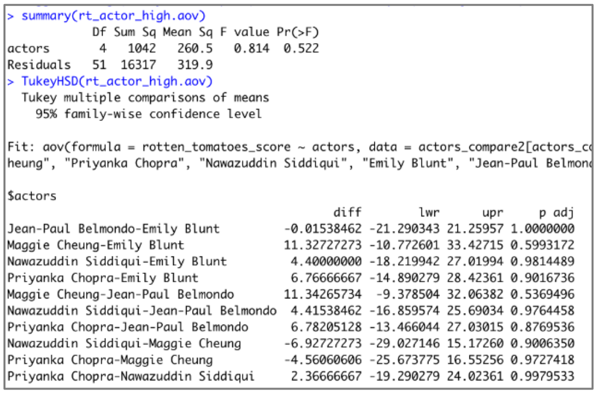


```
> summary(imdb_dir_high.aov)
            Df Sum Sq Mean Sq F value Pr(>F)
director     4  2.349  0.5871   1.532  0.203
Residuals   67 25.680  0.3833
> TukeyHSD(imdb_dir_high.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = imdb_score ~ director, data = dir_compare2[dir_compare2$direc
tin Scorsese", "Charles Chaplin", "Steven Spielberg", "François Truffaut"), ])

$director
                                        diff        lwr       upr     p adj
François Truffaut-Charles Chaplin  -0.1727273 -0.9127986 0.5673440 0.9652494
Hayao Miyazaki-Charles Chaplin      0.3803030 -0.3441861 1.1047922 0.5842461
Martin Scorsese-Charles Chaplin     0.2064935 -0.4928082 0.9057952 0.9210823
Steven Spielberg-Charles Chaplin   -0.0280303 -0.6599865 0.6039259 0.9999446
Hayao Miyazaki-François Truffaut    0.5530303 -0.1714588 1.2775194 0.2156520
Martin Scorsese-François Truffaut   0.3792208 -0.3200809 1.0785224 0.5531825
Steven Spielberg-François Truffaut  0.1446970 -0.4872592 0.7766532 0.9675643
Martin Scorsese-Hayao Miyazaki     -0.1738095 -0.8565992 0.5089802 0.9526761
Steven Spielberg-Hayao Miyazaki    -0.4083333 -1.0219681 0.2053014 0.3457967
Steven Spielberg-Martin Scorsese   -0.2345238 -0.8182069 0.3491592 0.7920342
```

Figure: IMDb Score vs Directors

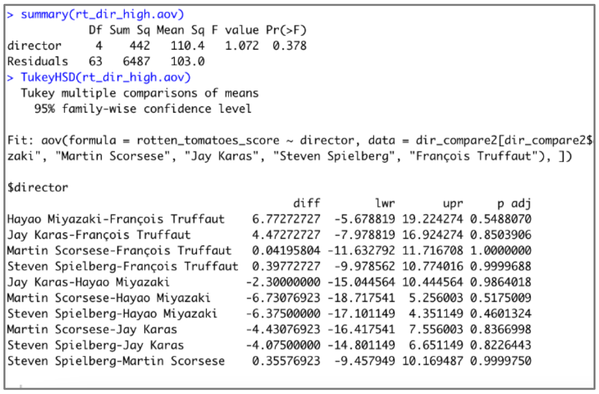


```
> summary(rt_actor_high.aov)
            Df Sum Sq Mean Sq F value Pr(>F)
actors       4   1042   260.5   0.814  0.522
Residuals   51  16317   319.9
> TukeyHSD(rt_actor_high.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = rotten_tomatoes_score ~ actors, data = actors_compare2[actors_c
heung", "Priyanka Chopra", "Nawazuddin Siddiqui", "Emily Blunt", "Jean-Paul Belmond

$actors
                                           diff        lwr      upr     p adj
Jean-Paul Belmondo-Emily Blunt          -0.01538462 -21.290343 21.25957 1.0000000
Maggie Cheung-Emily Blunt               11.32727273 -10.772601 33.42715 0.5993172
Nawazuddin Siddiqui-Emily Blunt          4.40000000 -18.219942 27.01994 0.9814489
Priyanka Chopra-Emily Blunt              6.76666667 -14.890279 28.42361 0.9016736
Maggie Cheung-Jean-Paul Belmondo        11.34265734  -9.378504 32.06382 0.5369496
Nawazuddin Siddiqui-Jean-Paul Belmondo   4.41538462 -16.859574 25.69034 0.9764458
Priyanka Chopra-Jean-Paul Belmondo       6.78205128 -13.466044 27.03015 0.8769536
Nawazuddin Siddiqui-Maggie Cheung       -6.92727273 -29.027146 15.17260 0.9006350
Priyanka Chopra-Maggie Cheung           -4.56060606 -25.673775 16.55256 0.9727418
Priyanka Chopra-Nawazuddin Siddiqui      2.36666667 -19.290279 24.02361 0.9979533
```

Figure: Rotten Tomatoes Score vs Language



```
> summary(rt_dir_high.aov)
            Df Sum Sq Mean Sq F value Pr(>F)
director     4    442   110.4   1.072  0.378
Residuals   63   6487   103.0
> TukeyHSD(rt_dir_high.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = rotten_tomatoes_score ~ director, data = dir_compare2[dir_compare2$
zaki", "Martin Scorsese", "Jay Karas", "Steven Spielberg", "François Truffaut"), ])

$director
                                         diff        lwr       upr     p adj
Hayao Miyazaki-François Truffaut   6.77272727  -5.678819 19.224274 0.5488070
Jay Karas-François Truffaut        4.47272727  -7.978819 16.924274 0.8503906
Martin Scorsese-François Truffaut  0.04195804 -11.632792 11.716708 1.0000000
Steven Spielberg-François Truffaut 0.39772727  -9.978562 10.774016 0.9999688
Jay Karas-Hayao Miyazaki          -2.30000000 -15.044564 10.444564 0.9864018
Martin Scorsese-Hayao Miyazaki    -6.73076923 -18.717541  5.256003 0.5175009
Steven Spielberg-Hayao Miyazaki   -6.37500000 -17.101149  4.351149 0.4601324
Martin Scorsese-Jay Karas         -4.43076923 -16.417541  7.556003 0.8366998
Steven Spielberg-Jay Karas        -4.07500000 -14.801149  6.651149 0.8226443
Steven Spielberg-Martin Scorsese   0.35576923  -9.457949 10.169487 0.9999750
```

Figure: Rotten Tomatoes Score vs Directors