Quantitative fundamentals

Using R

1. Obtained the data “Hollywoods Most Profitable Stories” on [public.tableau.com](https://public.tableau.com/en-us/s/resources?qt-overview_resources=1#qt-overview_resources), and we learn that this data is generated based on the dataset “Hollywood Budget” from InformationIsBeautiful.net.

2) The data includes a random sample of 74 movies released between 2007 to 2011, and genre, lead studio, audience score, rotten tomato score, profitability and worldwide gross.

**Genre:** There are 6 genres included in this dataset. Although some movies might be recognized with more than one genre, the genre recorded in this dataset is obtained from IMDB, a well-known online movie database.

**Lead studio:** The studio who mainly produce the movie. 12 studios are included, and some other are only classified as independent studio.

**Audience Score and Rotten Tomato Score:** The average audience rating for the movie and the average metascore of critic reviews respectively. These data are obtained from rotten tomato website. It is notable that there is a special “status” system of both audience score and critic reviews.

* **Audience Score:** Rotten(less than 60% positive), Fresh(at least 60% of reviews positive)
* **Critic reviews:** Rotten(less than 60% positive), Fresh(at least 60% of reviews positive) and Certified Fresh( 75% or higher).

**Worldwide gross:** The worldwide total revenue of the movie in million$.

**Profitability:** How much a movie can profit when compared to its budget.  
 (i.e. Profitability = Worldwide gross / Budget) Therefore,  
 If profitability < 1, the studio is losing money.

If profitability = 1, the studio is neither losing money or earning money.

If profitability > 1, the studio is earning money.

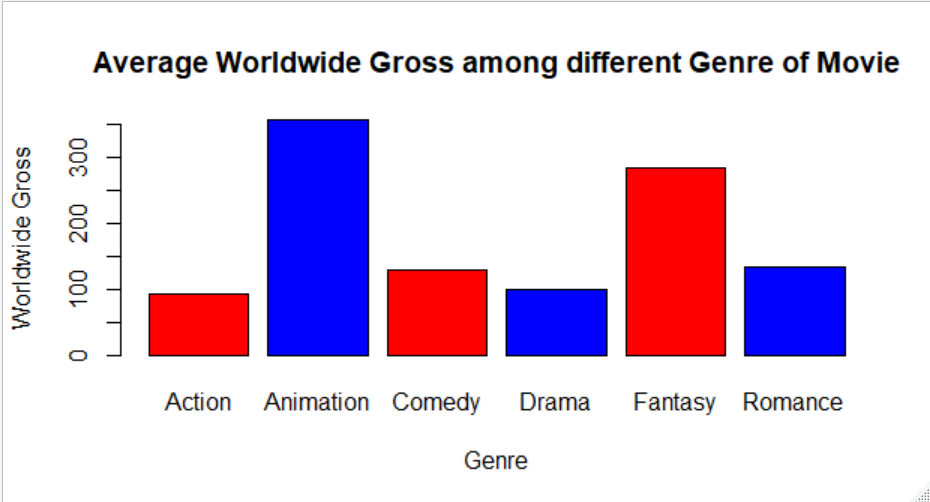
3)We design the study from the perspective of Film Studios. Since two of their main targets are to increase revenue and to create popular movies, so we decide to investigate factors related to worldwide gross and movie rating, including relationships between:

* Genre and Worldwide Gross/Profitability
* Audience score and Rotten Tomatoes score
* Audience score and Profitability/Worldwide Gross
* Audience score and Genre

4)

**QN1: Is there any relationship between genre and Worldwide Gross/profitability of the film?**

Plot a Bar Chart to see which genre of Movie seems to have largest means of worldwide gross profit:



Average worldwide gross is the largest for the Genre: Animation .i.e. we can say that on an average, the Animation Genre has been successful in terms of earning the highest gross over the years i. e from 2007 to 2011. However, the relationship of worldwide gross and genre need to be further evaluated.

**ANOVA**

Since some genre has relatively low observations. Therefore, we further group “Fantasy”, “Animation” and “Action” into a group called “Other. There are 5 observations for the group of “Other”

Null Hypothesis**: All the Population mean of worldwide gross are the same for each genre**

Alternate Hypothesis: **Not all population means of worldwide gross are equal for each genre**

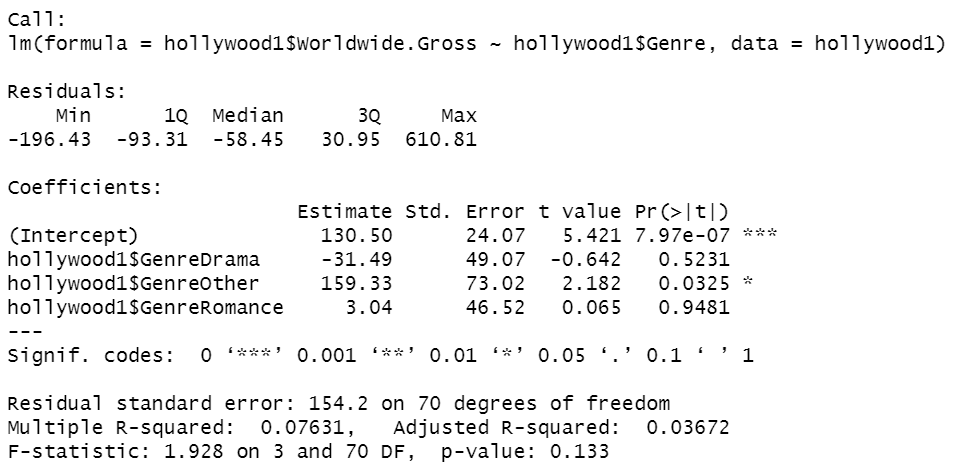
**P-value:**

The P value is 0.133 which is greater than 0.05. Therefore, we cannot reject the Null hypothesis. It implies that there is insufficient evidence to support that not all the population means are the same. Thus, we retain the belief that all the population mean of worldwide gross are the same for different genres.

**Test of Significance**

Null hypothesis: **There is no linear relationship between Worldwide Gross and Genre**

Alternate hypothesis: **There is a linear relationship between Worldwide Gross and Genre**

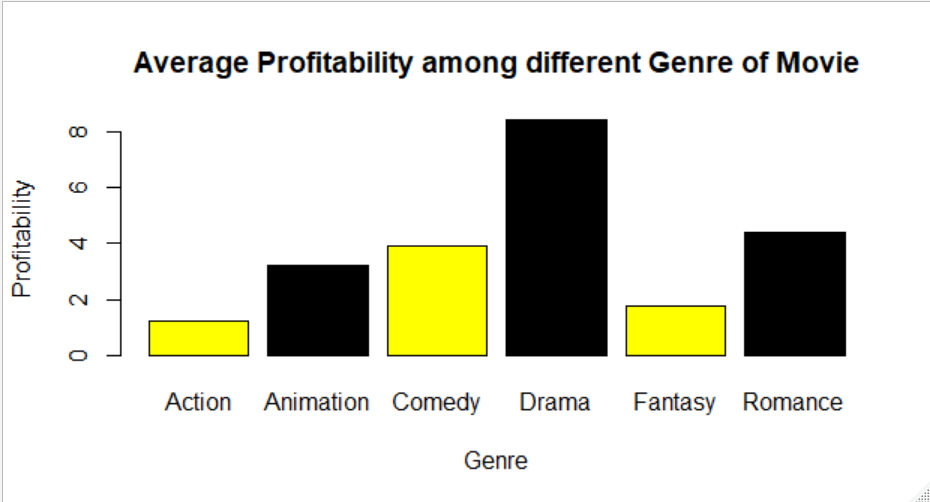
****

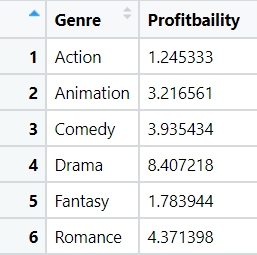
The P value of comedy and others group are less than 0.05(alpha) whether we can reject the null hypothesis. The P value of Drama and Romance are greater than alpha. Therefore, there is insufficient evidence to support that there is a linear relationship between worldwide gross and drama and romance movie However there are linear relationship between worldwide gross with drama and romance.

**Genre vs. Profitability**

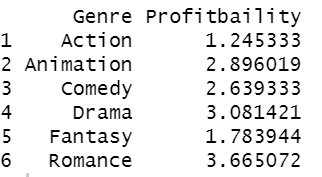
Apart from the column of worldwide gross, we can also evaluate how the column of profitability from the dataset interact with the genre.

Plot a bar chart to evaluate the mean profitability of different genre of movie in the dataset:





Median of Profitability on different genre of movie



As we explained earlier, profitability greater than 1 means that the movie can make a profit. If profitability is less than 1 means the movie is not able to make a profit. From the above findings, it seems that movies in Drama genre have the highest mean and median of profitability among genres of movies.

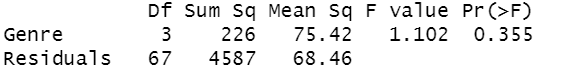
**ANOVA**

Similar to “Worldwide Gross”, we further group “Fantasy”, “Animation” and “Action” into a group called “Other. There are 5 observations for the group of “Other”.

Null Hypothesis**: All the Population mean of profitability are the same for each genre**

Alternate Hypothesis: **Not all population means of profitability are equal for each genre**

**P-value:**

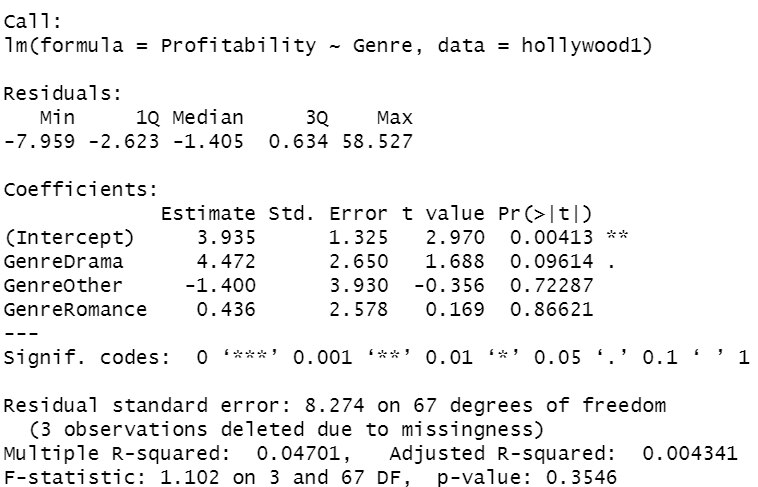
****

The p value is 0.355 which is greater than alpha. Therefore we are not able to reject the null hypothesis. Thus, there is insufficient evidence to support that all the population mean of worldwide.gross profit for are not the same for different genres. It implies that the population mean of profitability are equal for each genre.

**Test for Significance**

Ho: **There is no significant regression relationship between Audience score and Profitability**

Ha: **There is significant regression relationship between Audience score and Profitability**

****

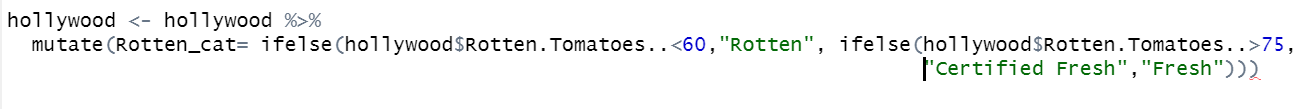
P value of all genres except comedy are greater than 0.05(alpha). Therefore, we are not able to reject the null hypothesis. It implies that there is no significant linear relationship between three groups of genre and profitability except comedy.

To conclude, worldwide gross and profitability has linear relationship with comedy movie. Worldwide gross has linear relationship with other group of movie(including Action, Fantasy and Animation). However, worldwide gross has no linear relationship with drama and romance. Thus profitability has no linear relationship with romance, drama and other genres.

**QN2**: **Is there any correlation between Audience Score and Rotten Tomato Score?**

We try to investigate if both scores are correlated and if the audience has the same opinion with film critics when they evaluate a movie.

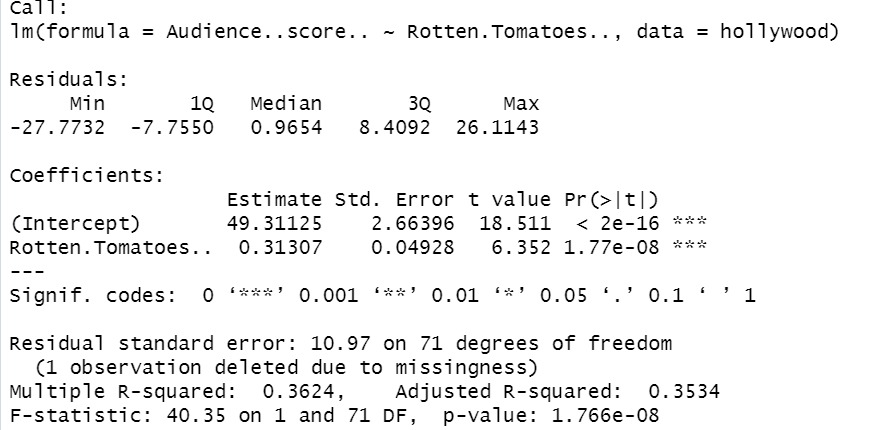
We can create a new column of Rotten Tomato score categories into to evaluate the correlation between audience score and rotten tomato score by the following code:



**Test of Significance**

**Null hypothesis:** There is no linear relationship between Rotten Tomato Score and Audience Score

**Alternate Hypothesis:** There is a linear relationship between Rotten Tomato Score and Audience Score

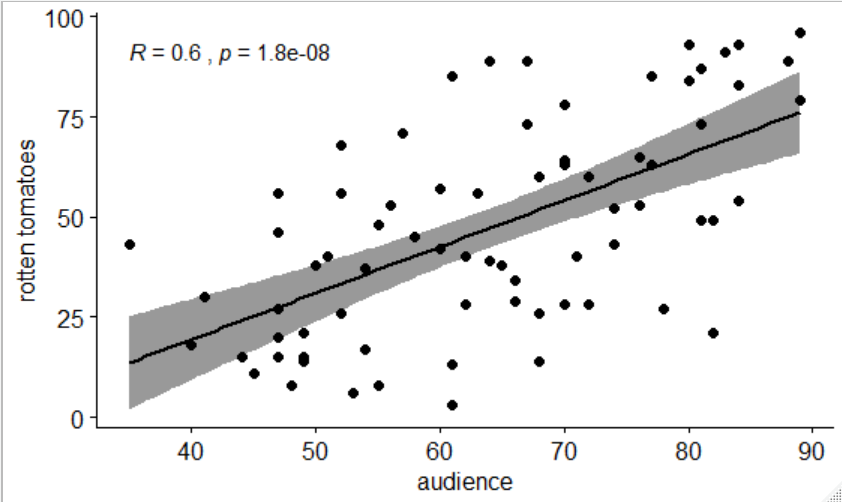
****

P value is 1.766e-08 which is less than alpha. Therefore, we are able to reject the Null hypothesis. There is sufficient evidence to support there is a linear relationship between audience score and rotten tomato score.

R-square: 0.362392

By calculating the r square. There is only around 36% of the variation of audience score are contributed by Rotten Tomato Score. Hence, there is not a good fit for the regression model.

Scatter Plot plotting Rotten Tomato Score Against Audience Score

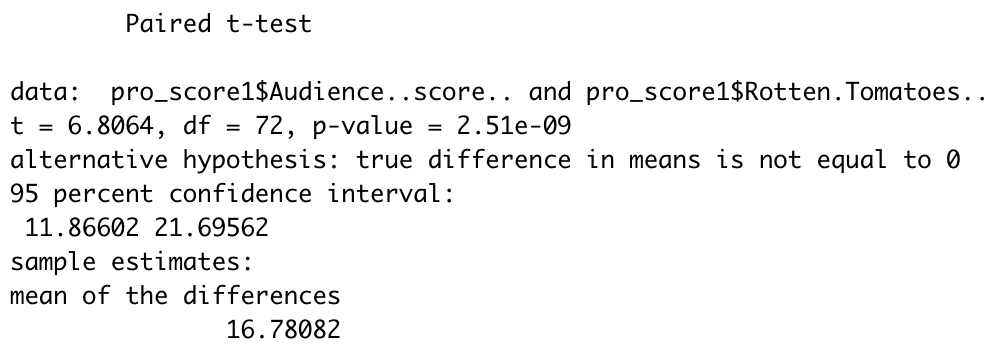
****

Paired test

Although the Audience Score can be explained by the Rotten Tomato Score, we still want to see if there is any difference for audiences and critics to rate a movie, so we perform a paired t-test to find out.

**Null hypothesis:** The mean difference between Rotten Tomato Score and Audience Score=0

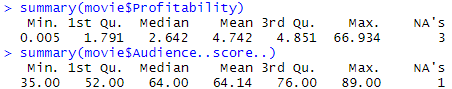
**Alternate Hypothesis:** The mean difference between Rotten Tomato Score and Audience Score is not equal to 0



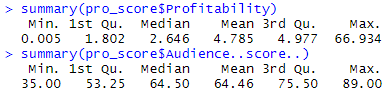
Since p-value <0.05, we can reject the null hypothesis and say difference occurs between audiences and critics, and on average, audiences tends to give 16.78 higher rating than critics. If we look into the 95% confidence interval, the difference could be from 11.87 to 21.70 scores.

**QN3: is there any relationship between the audience score and the profitability/worldwide gross?**

Before we go deep into the analysis, we can inspect the data first.



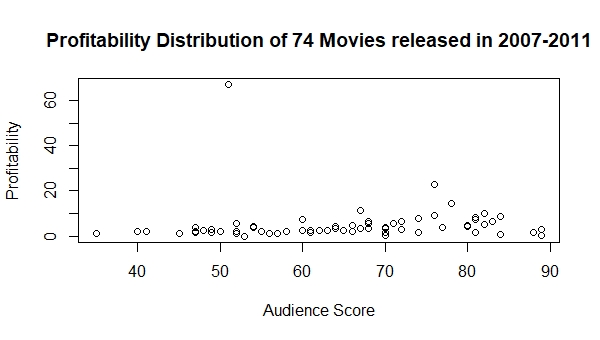
Since there are missing data, we will exclude them and inspect the data again.



As we can see, the range of profitability is very large, from 0.005 to 66.934, with the point estimate of 4.785, which means the movies produce profit 4.785 times more than their own budget on average. Also, we notice that the gap between 3rd Quantile and the Maximum is quite large, so outlier might exist in this case.

The mean and median of audience score are very close, only 0.04 difference between them. They are also close to the critical point, 60, to define the rotten/fresh status.

To evaluate if there is any relationship between audience score, we can first observe from a scatter plot.



According to the plot, most data are clustered between 0-300% profitability except one, so there is an outlier as mentioned before. Also, we can see the tendency that profitability is slowly increasing as audience score increase until audience score is around 80 or above.

We get the following results on calculating a correlation test between Audience Score and Profitability:

Pearson's product-moment correlation

data: data$Audience..score.. and data$Profitability

t = 0.34733, df = 68, p-value = 0.7294

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.1948179 0.2743438

sample estimates:

cor

0.04208256

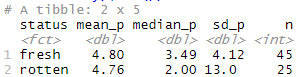
According to the above result,We compute the correlation coefficient = 0.04

Since its close to zero, we can say that there is no significant relationship between the Audience Score and Profitability.

The above graph confirms that there is no significant association between the Audience Score and Profitability.

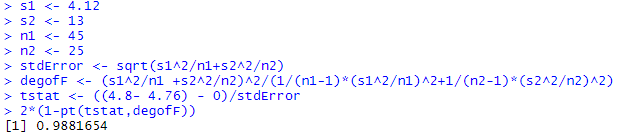
We would like to examine the relationship between these two variables by establishing new variable based on audience score and the Rotten Tomato status system, and thus it will divide data into 2 groups: Rotten and Fresh. Then, we can run hypothesis testing to see if there are any difference between the mean of profitability in two groups.

After assigning status to each movie data, there are 45 “fresh” and 25 “rotten” movies in the data. Although their mean profitability are close, the profitability are much more spread out for “rotten” movie since the standard deviation of “rotten” movies are larger than “fresh” movies.



Ho:**The difference of population means of profitability between “fresh” and “rotten” movie = 0**

Ha:**The difference of population means of profitability between “fresh” and “rotten” movie ≠ 0**

****

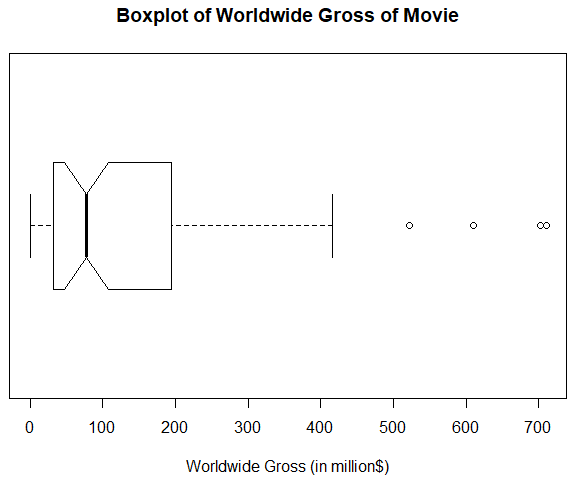
Since p-value >0.05, we do not have enough evidence to reject the null, so we cannot say there is a difference between population means of profitability of “fresh” and “rotten” movie.

Since profitability of a movie is highly affected by the budget, so we will try to investigate if there is any relationship between audience score and worldwide gross.

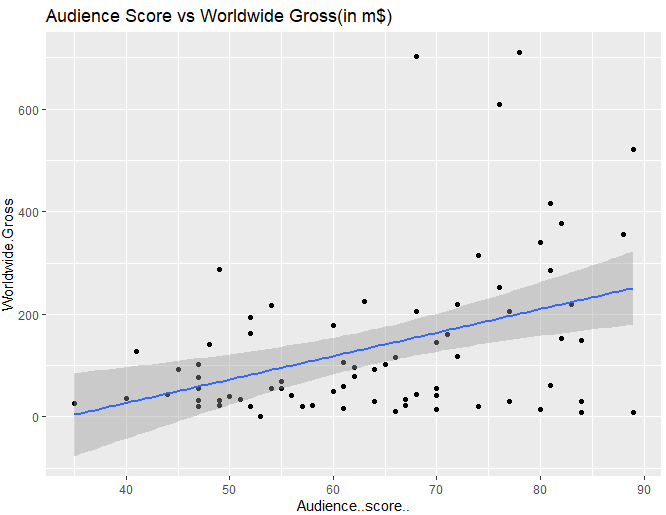
We can see the mean is larger than the median, so the distribution is right-skewed. Also, the data of 3rd quartile and maximum value has a huge gap, which indicates the possibility of existence of outlier.



The boxplot confirms the right-skewed distribution and existence of outliers.



Then, when we look into the scatter plot of Audience Score and worldwide gross, the graph shows a positive relationship.The Gross tends to increase as the Audience score increases.



**To check the nature of the relationship, the correlation is computed.**

data: data$Audience..score.. and data$Worldwide.Gross

t = 3.6268, df = 71, p-value = 0.0005366

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.1818282 0.5732792

sample estimates:

cor

0.3953566

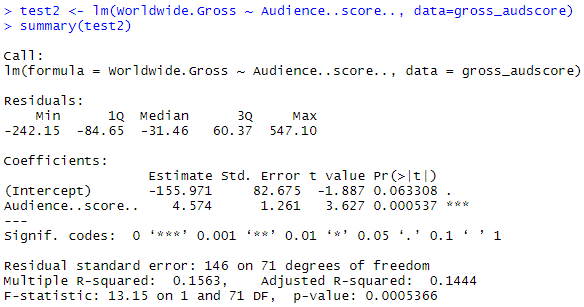
Here, we can see that the correlation coefficient is approx. = 0.40

Which confirms a positively natured relationship between the Audience Score and Worldwide Gross.

Test for significance

Ho: There is no significant regression relationship between Audience score and worldwide gross

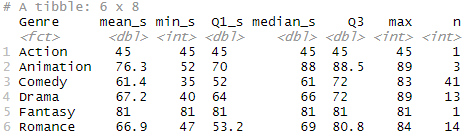
Ha: There is significant regression relationship between Audience score and worldwide gross

****

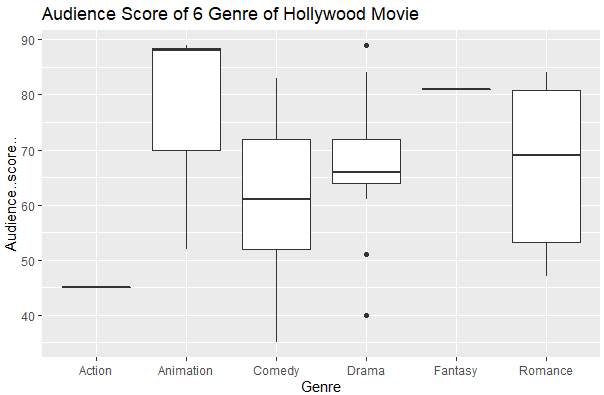
Since p-value < 0.05, we have sufficient evidence to reject the null hypothesis and conclude that the relationship between Audience score and Worldwide gross is significant. Although only 15.63% of the variability of worldwide gross of movie can be explained by the audience score, it actually makes sense in Hollywood because there are many other possible contributing factors such as number of advertisements, choice of actor/actress or number of awards winning.

**QN4:Is there any relationship between audience scores and genre of films?**

Now since we concluded that there is an increasing trend in gross as the audience ratings rise, it is interesting to find out if there is a typical inclination of audiences towards a particular genre.



The highest sample mean score belongs to the Fantasy genre with 81 mean score, when we compare the median, however, the Animation genre has the highest score, 88.

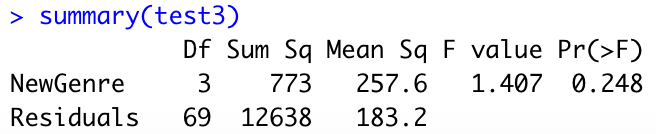


Also, since this dataset is collected by random sampling method, which means we cannot control how many movies in a genre can be included, so there is only one Action movie and one fantasy movie included. Therefore, the mean, minimum, Q1,median,Q3 and maximum are the same.

We then conduct ANOVA test to compare the mean audience score of different genres.

Ho: **The mean audience score of different genres of movies are the same.**

Ha: **Not all the mean audience score of different genres of movies are the same.**



Since the p-value>0.05, we do not have enough evidence to reject the null hypothesis, and we cannot prove that the mean audience score are different among different genres of movies.

**Conclusion/Suggestion:**

**Factors related to Profitability/Worldwide Gross**

There are some genres has linear relationship between genre and profitability and between genre and worldwide gross but not all of them. There are different reasons behind this. It may be due to profitability can be greatly influenced by the budget of the film. If the budget of the film is huge, it definitely decrease the profitability of the film. Therefore, genre may not necessarily the factor that can influence the profitability or worldwide gross profit. However, we can still evaluate the profitability and worldwide gross profit of different genre of movie from some of the leading studios by using descriptive statistics to give some references for movie maker to decide what genre of movie can be produced. Also, we are able to identify that worldwide gross of a movie can be explained by its audience score on rotten tomato website.

**Factors related to Audience/ Rotten Tomato Score**

We can conclude that the Audience score and the Rotten tomatoes score are in positive association with each other. However, only around 36% of the variation of audience score can be explained by Rotten tomatoes score. While it cannot be considered as a good fit, the paired test provides evidence that there is a difference between audiences and critics rating a movie, and on average, audiences tends to give 16.78 higher rating than critics.

For future film promotion strategy, we suggest that movie studio should monitor the audience rating on online movie website during the movie showing period. If the audience score fall below a critical point, marketing department should activate promotion strategies such as giving out discount for the next movie ticket if audience rate the movie online. In addition, studio could collaborate with youtube channel or key opinion leader specialized in film review area to produce positive viral contents for the movie.