IMDB Movie Data Analysis

Student Name:

Student Registration Number:

Institution:

**Table of Contents**

[1.0 Introduction 2](#_Toc120654626)

[1.1 Background Study 2](#_Toc120654627)

[1.2 Objectives 3](#_Toc120654628)

[1.3 Hypotheses 3](#_Toc120654629)

[2.0 Methods and Analysis 5](#_Toc120654630)

[2.1 Research Design 5](#_Toc120654631)

[2.2 Data 5](#_Toc120654632)

[2.3 Statistical Analysis Techniques 6](#_Toc120654633)

[3.0 Results and Discussion 7](#_Toc120654634)

[3.1 Descriptive Summary statistics 7](#_Toc120654635)

[3.2 Inferential Statistics 9](#_Toc120654636)

[3.2.1 Regression Analysis 9](#_Toc120654637)

[3.2.2 Independent Sample t-test 11](#_Toc120654638)

[3.3.3 One Way ANOVA 12](#_Toc120654639)

[3.3.4 Chi-Square Test (Cross Tabulation) 14](#_Toc120654640)

[4.0 Summary and Recommendations 15](#_Toc120654641)

[5.0 References 17](#_Toc120654642)

**List of Tables**

|  |  |
| --- | --- |
| Table 1 | Descriptive Summary Analysis |

**List of Figures**

|  |  |
| --- | --- |
| Figure 1 | Histogram plot for movie language |
| Figure 2 | Pie-Chart plot for movie color |
| Figure 3 | Regression Model Output |
| Figure 4 | Independent Sample t-test |
| Figure 5 | One-Way ANOVA Output |
| Figure 6 | Cross tabulation between movie color and movie language |

# 1.0 Introduction

## 1.1 Background Study

Data analytics plays a vital role in any organizational setting because it provides insights that help managerial decision makers or any given relevant stakeholder to make well-informed and appropriate decisions (Sigler et al. 2020, pp. 511-514). For instance, analysis of customer satisfaction survey data will help an organization to know which factors are important at influencing customer satisfactions. In this case study, IMDB (Internet Movie DataBase) Movie data will be analyzed in IBM SPSS 26 in order to get some insights from the data set. IMDB is the most authoritative and popular source for movies, celebrities and Television shows content (Dhir and Raj, 2018, pp. 385-390). It is an online database providing information about video games, TV series, home videos, movies as well as online streaming of content. It also provides information about the casts of the particular movie or TV series or video, the crew involved in production together with their individual biographies, ratings, trivia, plot summaries as well as fan and critical reviews (Dhir and Raj, 2018, pp. 385-390). Through this database, consumers of the movie product get to engage with each other and get an understanding of the movie and from reviews of other customers, people can be encouraged to watch a particular movie. One can also get encouraged to watch a particular movie after finding out that one of his or her favorite actor or actress, producer, director is among the cast. As a result, the IMDB database is significant to the growth and productivity of the movie industry.

In this case study, the IMDB Movie data set will be analyzed to provide several insights from the movie at least to get an understanding of what is happening. The objectives of this study are as follows;

## 1.2 Objectives

1. To determine if gross income of a movie can be explained or predicted by movie budget, duration of movie, average Facebook actor’s likes, total cast Facebook likes, number of voted users, IMDB Score and Aspect ratio (Regression Analysis).
2. To perform an Independent Sample t-test to determine if IMDB movie score, gross income and budget attributes have significant difference between color and black white movies.
3. To perform ANOVA test to determine if IMDB movie score, gross income and budget attributes have significant difference among IMDB new
4. To perform a cross tabulation analysis (Chi-Square test) to determine if there is significant association between movie color and movie language.

1.3 Hypotheses.

1.

Ho: Movie budget insignificantly predict gross income of a movie.

H1: Movie budget significantly predict gross income of a movie.

2.

Ho: Duration of movie insignificantly predict gross income of a movie

H1: Duration of movie significantly predict gross income of a movie

3.

Ho: Average Facebook actor’s likes insignificantly predict gross income of a movie

H1: Average Facebook actor’s likes significantly predict gross income of a movie

4.

Ho: Total cast Facebook likes insignificantly predict gross income of a movie

H1: Total cast Facebook likes Significantly predict gross income of a movie

5.

Ho: Number of voted users insignificantly predict gross income of a movie

H1: Number of voted users significantly predict gross income of a movie

6.

Ho: IMDB Score insignificantly predict gross income of a movie

H1: Ho: IMDB Score significantly predict gross income of a movie

7.

Ho: Aspect ratio insignificantly predict gross income of a movie.

H1: Aspect ratio significantly predict gross income of a movie

8

Ho: There is no significance difference between movie color and IMDB movie score

H1: There is significance difference between movie color and IMDB movie score

9.

Ho: There is no significance difference between movie color and budget.

H1: There is significance difference between movie color and budget

10.

Ho: There is no significance difference between movie color and gross income

H1: There is significance difference between movie color and gross income.

11.

Ho: There is no significance difference between IMDB new and IMDB movie score

H1: There is significance difference between IMDB new and IMDB movie score

12.

Ho: There is no significance difference between IMDB new and gross income

H1: There is significance difference between IMDB new and gross income

13.

Ho: There is no significance difference between IMDB new and budget

H1: There is significance difference between IMDB new and budget

14.

Ho: There is no significance association between movie colors and among movie languages.

H1: There is significance association between movie colors and movie languages.

# 2.0 Methods and Analysis

## 2.1 Research Design

The three research designs mostly used are qualitative, quantitative and mixed method designs. Qualitative is whereby data that is no numeric is used in the analysis while quantitative is whereby numerical data is employed in the analysis (Rutberg and Bouikidis, 2018, pp. 200-213). Mixed method design applies both qualitative and quantitative. In this case study, quantitative research design was used because the study entailed analysing numerical data using IBM SPSS 26 software.

## 2.2 Data

In any given research, data source is either primary or secondary. Secondary source of data is a pre-collected or pre-recorded data that has been collected, stored and ready to be used for analysis while on the other hand, primary data source is one whereby the researchers collect data by themselves for instance through surveys, interviews (Cole et al. 2018, pp. 165-173). In this case study, secondary data set called IMDB Movie Data was used for analysis.

IMDB Movie Data had 5043 observations. The data had to be cleaned by dropping any missing observations or observations with entries FALSE after which the sample size was reduced to 3726. For categorical variable color/movie color ‘Color’ was replaced with 1 while “Black and White” was replaced with 2. After importing the cleaned data set in SPSS, the variables actor\_1\_facebook\_likes, actor\_2\_facebook\_likes and actor\_3\_facebook\_likes were transformed into average Facebook actors’ likes variable using TRANSFORMATION function in SPSS.

## 2.3 Statistical Analysis Techniques

Various statistical techniques were performed on the data set and the first one was descriptive summary analysis. Descriptive summary analysis was performed on the continuous variables whereby mean, standard deviation, skewness, kurtosis and Shapiro-Wilk was done. Categorical variables were explored through histogram plots and pie-chart. Then inferential statistics were performed in which the first one was multiple linear regression analysis and it was performed to determine objective and hypothesis 1 of the study. Regression analysis is used to determine impact of factor or predictor or independent variable on a dependent variable. In our case study, dependent variable was gross income while independent variables were Movie budget, duration of movie, average Facebook actor’s likes, total cast Facebook likes, number of voted users, IMDB Score and Aspect ratio. Another inferential statistics was independent sample t-test. It is used to determine if the difference between two independent groups is significant. In this case study, the test was used to determine if IMDB movie score, gross income and budget attributes have significant difference between color and black white movies, for objective and hypothesis 2. Analysis of Variance (ANOVA) was also used and this test is used to determine if the difference of more than two groups is significant or not. The test was used to determine if IMDB movie score, gross income and budget attributes have significant difference among movie language, for objective and hypothesis 3. The last inferential statistics used was Chi-Square test (cross tabulation) which is used to determine the degree of correlation in nominal variables. In this case study, the test was used to determine if there is significant different in movie color as well as movie language, for objective and hypothesis 4.

# 3.0 Results and Discussion

## 3.1 Descriptive Summary statistics

*Table 1: Descriptive summary statistics*

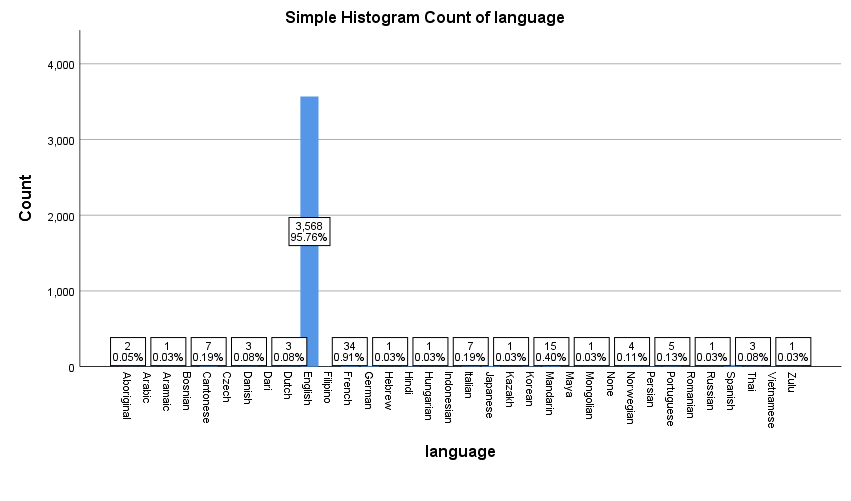
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Continuous Variables | Mean | Std.Dev | Skewness | Kurtosis | Shapiro-Wilk Test (p) |
| Gross income | 52791315.55 | 70539969.37 | 3.02 | 13.86 | 0.00 |
| Number of voted users | 106441.41 | 152466.62 | 3.64 | 19.85 | 0.00 |
| Budget | 46388179.61 | 226898086.00 | 44.01 | 2260.92 | 0.00 |
| IMDB Score | 6.48 | 1.05 | -0.75 | 1.25 | 0.00 |
| Aspect ratio | 2.11 | 0.35 | 16.09 | 638.36 | 0.00 |
| Average actor’s Facebook likes | 3531.36 | 6207.02 | 13.83 | 411.49 | 0.00 |
| Duration | 110.34 | 22.68 | 2.40 | 12.61 | 0.00 |
| Cast total Facebook likes | 11576.80 | 19180.82 | 12.87 | 367.81 | 0.00 |

The variables gross income, number of voted users, budget, average actor’s Facebook likes, duration and cast total Facebook likes had significantly large standard deviations in relation to their respective mean values. This means that there was high degree of variation or fluctuation in the measurements of these variables from their mean values (Mishra et al. 2019, p. 67). The variables IMDB score and aspect ratio can be seen to have insignificant standard deviations in relation to their mean values. This shows these variables had no high degree of fluctuation.

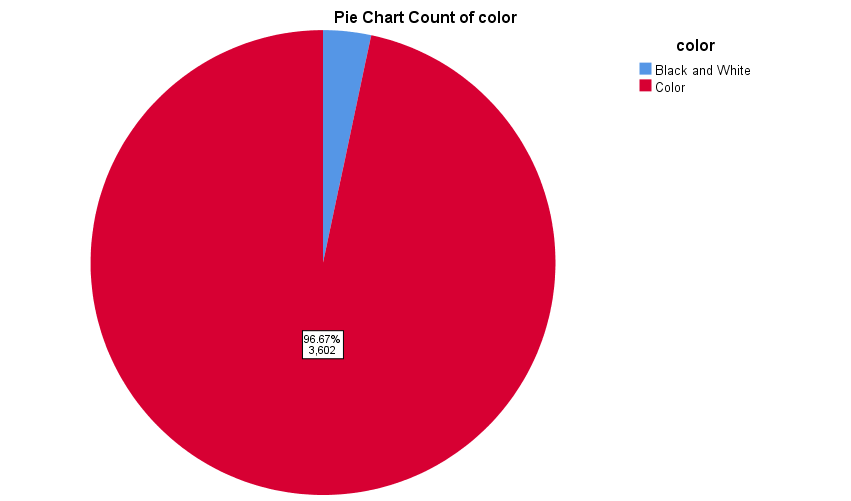
To explain the shape of distribution, skewness and kurtosis were used. For the skewness, the variables gross income, number of voted users, budget, aspect ratio, average actor’s Facebook likes, duration and cast total had positive value indicating their distributions were skewed to the right while for the IMDB score, its distribution was skewed to the left because it had negative value (Cain et al. 2017, pp. 1716-1735). All of the kurtosis values were positive meaning that the peak of these variables were not too flat (Cain et al. 2017, pp. 1716-1735). Shapiro-Wilk was used to test for normality and since p<0.05 for all variables, then their distributions were normal.

For categorical variables, the following is the histogram and pie-chart plots.

*Figure 1: Histogram plot for movie language*



*Figure 2: Pie- Chart plot for movie color*



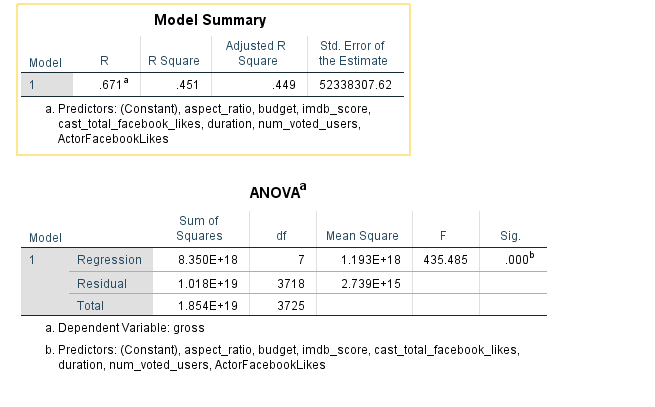
From figure 1, most movies are in English language, 95.76% of the movies while the least languages are Zulu, Russian, Mongolian, Hebrew, Aromaic, Kazaih at 0.03% of the movies. In figure 2, most of the movies are colored (96.67%) while only 3.33% are in black and white.

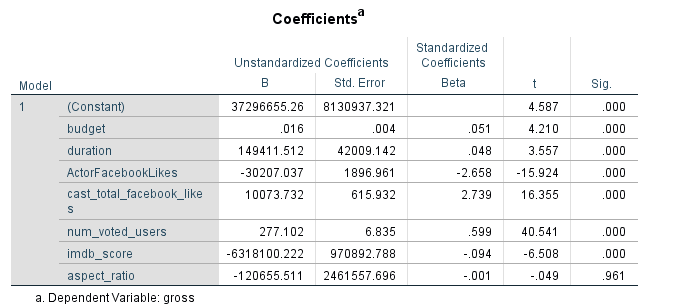
## 3.2 Inferential Statistics

### 3.2.1 Regression Analysis

This was used to determine objective 1 and hypotheses 1, 2, 3, 4, 5, 6 and 7.

*Figure 3: Regression Model Output*





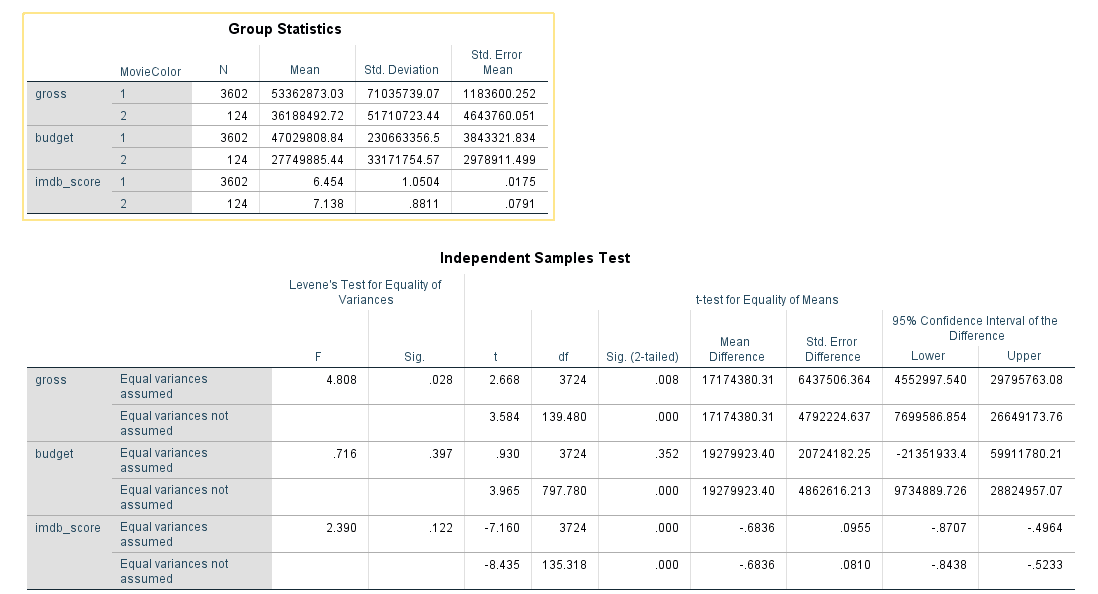
From figure 3, R-Squared value was 45.10% implying that the only 45.10% variability of the dependent variable were explained by the model (Sarstedt and Mooi, 2019, pp. 209-256). This shows the model was averagely good. F (7, 3718) = 435.50, p<0.05 and since p<0.05, the model was generally good in fitting the data set (Sarstedt and Mooi, 2019, pp. 209-256).

For the independent variables’ coefficients, budget (β=0.05, p<0.05), duration (β=0.05, p<0.05), average actor’s Facebook likes (β=-2.66, p<0.05), cast total Facebook likes (β=2.74, p<0.05), number of voted users (β=0.60, p<0.05), IMDB score (β=-0.09, p<0.05) and aspect ratio (β=-0.001, p>0.05). We can see that movie budget, duration, cast total Facebook like and number of voted users are positively related to the gross income of a movie hence an increase in either of these variables will lead to an increase in the gross income of a movie (Sarstedt and Mooi, 2019, pp. 209-256). On the other hand, the variables average actors’ Facebook likes, IMDB score and aspect ratio score are negatively related to gross income of a movie hence an increase in either of these variables will lead to a decrease in gross income of a movie. For all of variables except aspect ratio, p<0.05 hence these variables are significant at explaining, predicting or determining gross income of a movie (Sarstedt and Mooi, 2019, pp. 209-256). Only aspect ratio is not significant. Therefore, Ho in Hypothesis 1, 2, 3, 4, 5 and 6 were rejected and H1 concluded to be true and hence, Movie budget, duration of movie, average Facebook actor’s likes, total cast Facebook likes, number of voted users, IMDB Score significantly predict gross income of a movie. For hypothesis 7, H1 is true since p>0.05 hence aspect ratio does not significantly predict movie income.

### 3.2.2 Independent Sample t-test

This was used to determine objective 2 and hypotheses 8, 9 and 10. Below is the output.

*Figure 4: Independent Sample t-test*

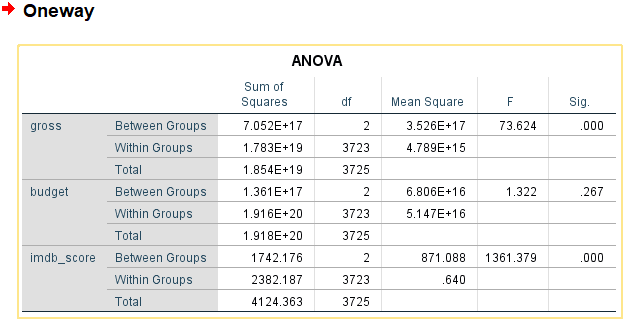


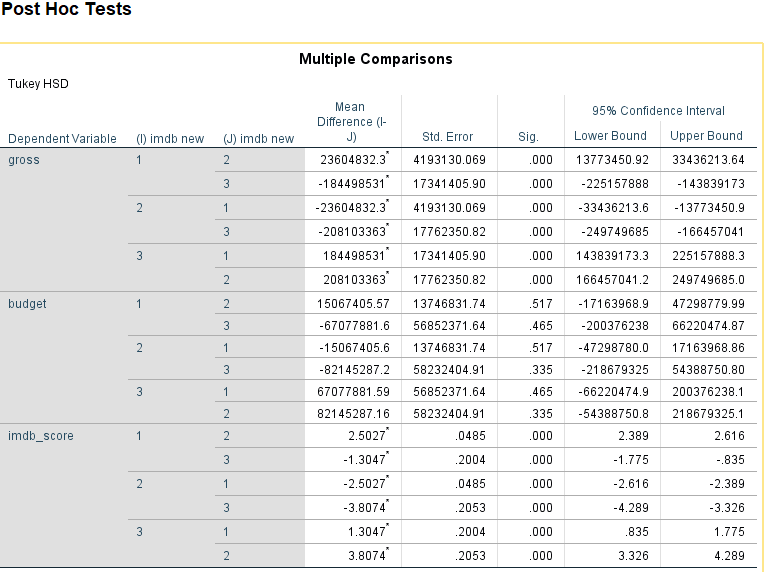
For the variable gross, p= 0.03 hence p<0.05. We reject Ho and accept H1 to be true and hence conclude that there is a significant difference in gross income when a movie is either colored or black and white hence H1 for Hypothesis 10 is true. For the variables budget and IMDB score, p= 0.40 and 0.12 respectively hence p>0.05. We fail to reject Ho for both variables and conclude that when it comes to the movie budget, there is no significance difference if either the movie is colored or black and white and IMDB score, there is no significance difference in the score in either colored or black and white movie. Hence Ho for hypothesis 8 and 9 are true.

### 3.3.3 One Way ANOVA

This was used to determine objective 3 and hypotheses 11, 12 and 13.

*Figure 5: One-Way ANOVE output*



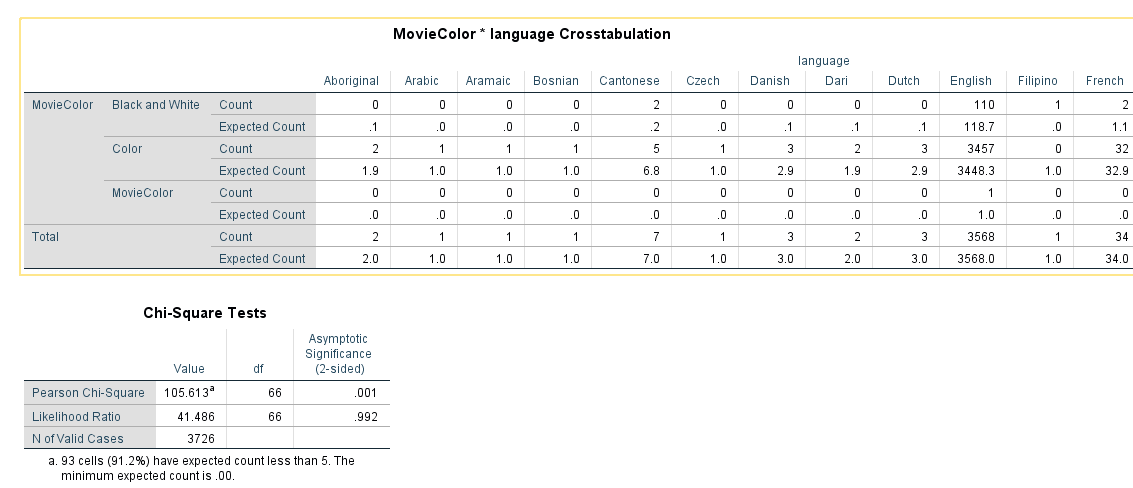


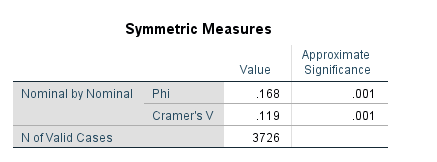
The variable IMDB new had three groups namely 1= new, 2= Not new and 3= False. From figure 5, the variables gross income and IMDB score have p<0.05 hence rejecting the null hypothesis and concluding that there is significance difference of gross income and IMDB score among the three groups of IMDB new, hence for hypothesis 11 and 12, H1 is true. For the variable budget, p-value = 0.27, hence p>0.05 thus we fail to reject null hypothesis and conclude it to be true hence for hypothesis 13, Ho is true. Therefore, there is no significance difference in the movie budget in relation to whether the movie new, old or False on the IMDB online database. For the Post Hoc test result, it is used to determine how each group (No, Yes, FALSE) in the IMDB new is different from each other in each of the various variables namely gross income, budget and IMDB score. For the variable gross and IMDB score, it can be seen that all of the three independent groups are significantly different (p<0.05) while for the variable budget, it can be seen that all of the three independent groups of IMDB are not significantly different (p>0.05).

### 3.3.4 Chi-Square Test (Cross Tabulation)

Used to determine the degree of relationship between nominal variables (Lin et al. 2015, pp. 438-458). It is the opposite for Pearson’s correlation analysis that is used to determine the strength of association between two numerical variables. In this case study, the test was used to determine if there was a significance different between movie color and movie languages (Hypothesis 14).

*Figure 6: Cross tabulation between movie color and movie language*





The expected count shows the expected value when there is association between the variables. For instance, from figure 6, for black and white movie, the expected observation is 1 for aboriginal movie while when the movie is colored, it is expected that they are 1.9 of such movies in the Aboriginal language. The rest applies to all of the languages. Now, the chi-square test is used to determine if the difference between the expected count and count is significant enough for there to be a significant association between the two groups namely movie color and movie language. The Pearson Chi-Square test statistic is 105.61 while the Asymptotic Significance which is our p-value is 0.001. Clearly, p<0.05 implying significance difference hence we reject the null hypothesis in hypothesis 14 (Lin et al. 2015, pp. 438-458). Therefore, we conclude that there is a significant relationship between movie color and movie language.

# 4.0 Summary and Recommendations

Data analytics is really important in providing insights that help to make appropriate organizational decisions. This case study was aimed at analyzing IMDB Movie Data in SPSS 26. The analytics techniques used were descriptive summary analysis, multiple linear regression, independent sample t-test and Chi-Square test. Descriptive statistics results indicated that the variables gross income, number of voted users, budget, average actor’s Facebook likes, duration and cast total Facebook likes had high degree of variation or fluctuation while the variables IMDB score and aspect ratio had no high degree of fluctuation. The Shapiro-Wilk test indicated that all of these variables were normally distributed. For the multiple linear regression results, movie budget, duration, cast total Facebook like and number of voted users are positively related to the gross income while actors’ Facebook likes, IMDB score and aspect ratio score were negative related to gross income of the movie. The results also indicated that except aspect ratio, the rest of the variables were significant at explaining or predicting movie gross income hence these variables can be reliable when making projections of a movie. Independent sample t-test was then used to determine if there is significance difference in movie score, gross income and budget attributes when a movie is colored or is black and white. The results indicated that for gross income variable, it was significantly different for colored and black and white movies. There was no significance difference in budget and IMDB score. One-Way ANOVA results showed that the three groups in IMDB New were significantly different for the variables gross income and IMBB score and on the other hand, they were not significant for the variable movie budget. The Chi-Square test was also performed and it indicated that there is a strong association between movie color and movie language.

Since the regression results indicated that there is a significant relationship between movie budget, movie income and duration of the movie, I would recommend movie producers, relevant stakeholders in the project to pay much attention on these variables when making movies so that to achieve enough and even high revenues while keeping the movie fans entertained and satisfied.

# 5.0 References

Cain, M.K., Zhang, Z. and Yuan, K.H., 2017. Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior research methods*, *49*(5), pp.1716-1735. <https://link.springer.com/article/10.3758/s13428-016-0814-1>.

Cole, A.P., Friedlander, D.F. and Trinh, Q.D., 2018, April. Secondary data sources for health services research in urologic oncology. In *Urologic Oncology: Seminars and Original Investigations* (Vol. 36, No. 4, pp. 165-173). Elsevier. <https://doi.org/10.1016/j.urolonc.2017.08.008>.

Dhir, R. and Raj, A., 2018, December. Movie success prediction using machine learning algorithms and their comparison. In *2018 first international conference on secure cyber computing and communication (ICSCCC)* (pp. 385-390). IEEE.

Lin, J.J., Chang, C.H. and Pal, N., 2015. A revisit to contingency table and tests of independence: bootstrap is preferred to Chi-square approximations as well as Fisher’s exact test. *Journal of biopharmaceutical statistics*, *25*(3), pp.438-458. <https://www.tandfonline.com/doi/abs/10.1080/10543406.2014.920851>.

Mishra, P., Pandey, C.M., Singh, U., Gupta, A., Sahu, C. and Keshri, A., 2019. Descriptive statistics and normality tests for statistical data. *Annals of cardiac anaesthesia*, *22*(1), p.67. <https://doi.org/10.4103%2Faca.ACA_157_18>.

Rutberg, S. and Bouikidis, C.D., 2018. Focusing on the fundamentals: A simplistic differentiation between qualitative and quantitative research. *Nephrology Nursing Journal*, *45*(2), pp.209-213. <https://www.proquest.com/openview/af62fd5b0442e59b2729d9fcf7348456/1?pq-origsite=gscholar&cbl=45638>.

Sarstedt, M. and Mooi, E., 2019. Regression analysis. In *A Concise Guide to Market Research* (pp. 209-256). Springer, Berlin, Heidelberg. <https://link.springer.com/chapter/10.1007/978-3-662-56707-4_7>.

Sigler, R., Morrison, J. and Moriarity, A.K., 2020. The importance of data analytics and business intelligence for radiologists. *Journal of the American College of Radiology*, *17*(4), pp.511-514. <https://www.jacr.org/article/S1546-1440(19)31485-1/fulltext>.