



COMPONENT 2

COURSEWORK: INDIVIDUAL REPORT

MODULE TITLE: DATA HANDLING & BUSINESS INTELLIGENCE

MODULE CODE: LCBB5014

STUDENT NAME: CRINA DENISA GHIGA

STUDENT ID: 2322635

LECTURER'S NAME: XAVIER MATIENI

UNIVERSITY OF WALES TRINITY SAINT DAVID

Table of Contents

1. Introduction	3
2. Data cleaning and Data preparation	3
3. Descriptive Statistics and Summary	4
4. Trend Analysis by Data Visualisation	5
5. Statistical analysis in SPSS	10
5.1. Hypothesis testing through an independent t-test	10
5.2. Correlation Analysis	11
5.3. Movie description relationship with rating and language influence over engagement	12
6. Insights and Recommendations	13
References	15

1. Introduction

The Letterbox Movie Classification Dataset will be examined in this report to determine the major factors that have the greatest influence on popular films and the ratings given to the films by the audience. The use of data to make decisions is a crucial concept in the modern media and entertainment sector to comprehend the choice of the viewers, enhance the process of creating the content and enhance their communication with the audience. The purpose of this research is to determine user behaviour trends and quantify the predictors of the success of films with the help of business intelligence methods.

A combination of Microsoft Excel and SPSS software was utilised for the realisation of the objectives of this paper. Data cleaning, data preparation, and data visual trend analysis were carried out in Excel software, while statistical analysis was done in SPSS, such as descriptive statistics, correlation analysis, and regression modelling. The data that will be used for the analysis of this paper includes 10,000 movie data points, and the variables include average rating, likes, watched, number of fans, and runtime. The above-mentioned variables will also provide information regarding the interests and quality of content of the audience.

The key objectives of the study include reviewing the pattern of viewer behaviour, identifying the correlation among engagement measures and rating, and determining the degree of impact of likes, fans, and runtime as independent variables on average film ratings. Lastly, evidence-based policy measures will be proposed based on the findings, which will enhance the performance of the firms, operational effectiveness, and satisfaction of the viewers within the film industry.

2. Data cleaning and Data preparation

This was done in Microsoft Excel to clean and prepare the data to render it accurate, uniform and fit to analyse it. The data was first analysed to retrieve missing and conflicting data of the key variables (runtime, likes, watches, fans and average rating). Records containing critical missing values were also checked; in the case where the small holes were not critical, such as the fact that nothing was found about engagement, it was removed to ensure data integrity, as imputation may cause the variables that are related to engagement to be biased.

Moreover, there was data validation for checking duplication or any possible errors related to formatting. Data normalisation was carried out in such a way that all the variable values could be categorised easily. Moreover, the chances of ambiguity in analysis were kept to a minimum level. For making sure that data was in the correct form and no non-numeric value or anomaly was present, numerical values (like number of likes, watches, fans) were validated. Also, outlier testing was conducted for the values of 'runtime', like high or low values of runtime.

The preprocessing tasks are essential in business intelligence since they enhance the reliability and validity of subsequent analyses (Mu, Cardelli and Ferrari, 2026). The statistical results are correct and significant because the data is clean and well-organised (Alexakis et al., 2025). It is possible to interpret data more easily and make effective decisions based on data in the film performance and the participation of audiences with the help of data cleaning.

3. Descriptive Statistics and Summary

Descriptive Statistics									
	N Statistic	Minimum Statistic	Maximum Statistic	Mean Statistic	Std. Deviation Statistic	Skewness		Kurtosis	
						Statistic	Std. Error	Statistic	Std. Error
Average_rating	10000	1.080000000	4.710000000	3.268535188	.5257438723	-.408	.024	.394	.049
Runtime	10000	2.000000	907.000000	103.1693515	38.07073527	5.650	.024	72.712	.049
Watches	10000	2	5195503	170314.82	458532.917	5.130	.024	32.539	.049
Fans	10000	0	326000	1652.55	9368.258	14.622	.024	324.808	.049
Valid N (listwise)	10000								

Figure 1: Descriptive statistics

The descriptive statistics of the data present underlying trends and patterns of the data. The results indicate that the mean rating of the movies is approximately 3.27 with a minimum rating of 1.08 and a maximum rating of 4.71. This implies that the bulk of the films are rated in the moderately positive range with a relatively small range (standard deviation ≈ 0.53), meaning that there is some consistency in audience scoring behaviour.

The variables in relation to engagement, like likes, watches, and fans, show a lot of variation. The mean value of likes is 43,734, while the minimum and maximum values are 0 and 2.7 million and above, showing that there is a large variation in the popularity of the movies. The same trend

applies to the value of watches, whose mean is about 170,315, with the largest value being above 5.1 million, showing that only a few films have many viewers. Similarly, the fans variable shows variation since the mean is about 1,653 while the maximum is 326,000.

The average run time is about 103 minutes with a minimum and maximum of 2 and 907 minutes, respectively. This generality implies that all sorts of content of this nature will be incorporated, such as short movies and long-term projects. The standard deviation is, however, relatively equal, which implies that most movies are concentrated around the average feature-length.

In general, the descriptive analysis shows that the audience rating is generally quite stable, but the degree of engagement is quite variable, and it can be assumed that the popularity of a film is concentrated in several films.

4. Trend Analysis by Data Visualisation

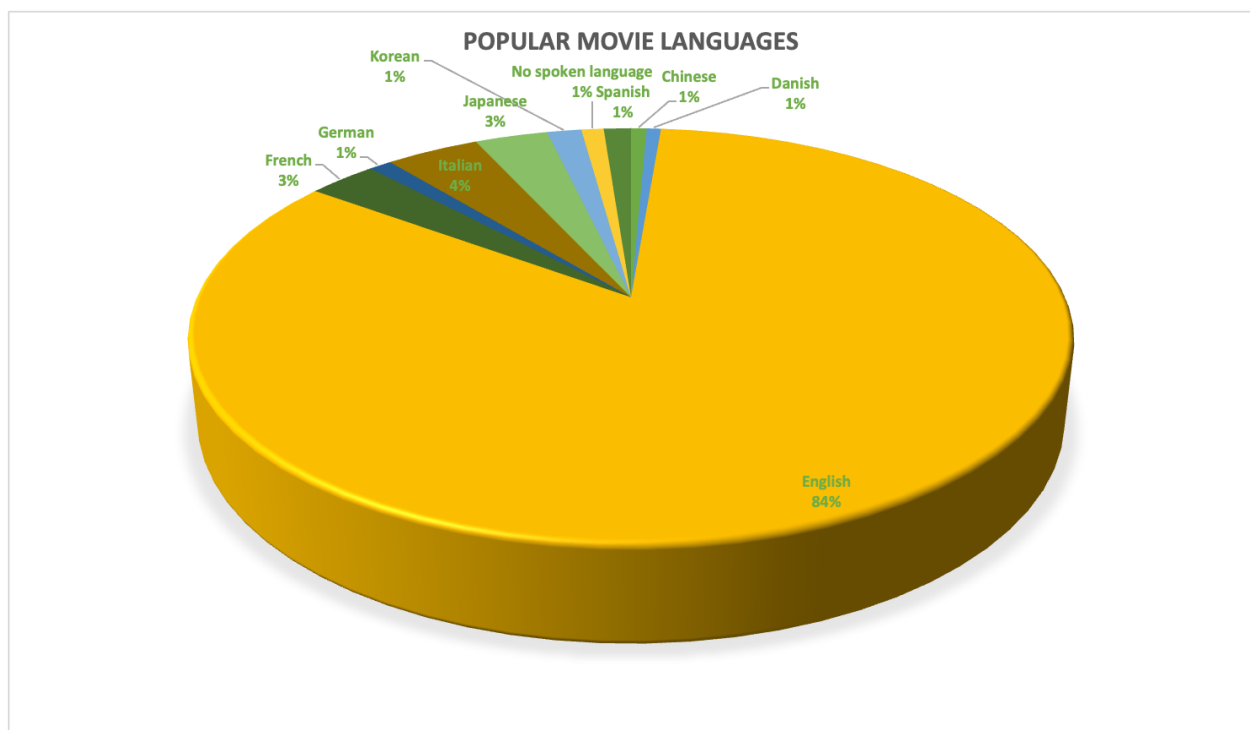


Figure 2: Popular Movie language

The distribution is skewed as indicated in Figure 2, with most of the data being constituted by English (84%). This implies that most of the movies are made or watched in English, which means that it has an influence on the film industry all around the world. The shares in other languages are far less, and they are Italian, 4, French and Japanese, 3, and German, 1. Overall, the chart shows that a deficit in linguistic diversity exists, which means that it can be enhanced by adding more non-English content to draw the interest of a broader international audience.

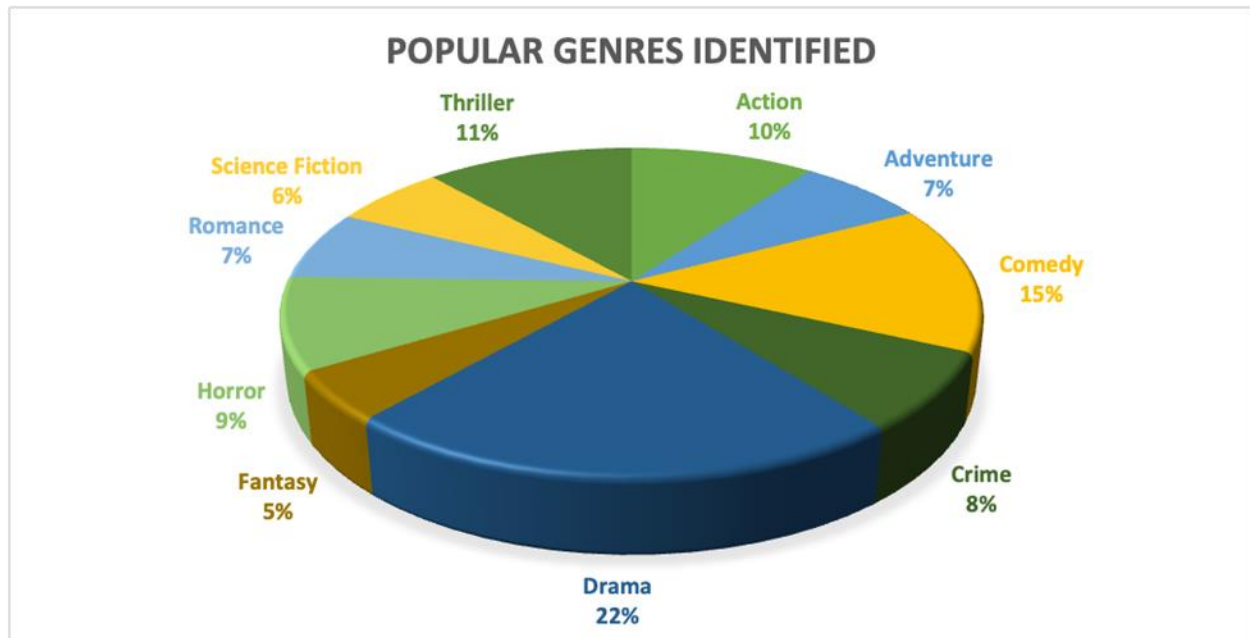


Figure 3: Popular Movie Genre

Figure 3 determines the distribution of genres of movies in the dataset. The most popular genre is drama with 22%. The one that comes next in popularity is the comedy at 15, meaning that it has a broad appeal by the viewers. Other salient shares include Thriller (11) and Action (10), which demonstrate that individuals are interested in content of high intensity and speed. In general, as the chart shows, the taste towards drama and entertainment-oriented genres is also evident, possibly affecting the production of content and marketing.

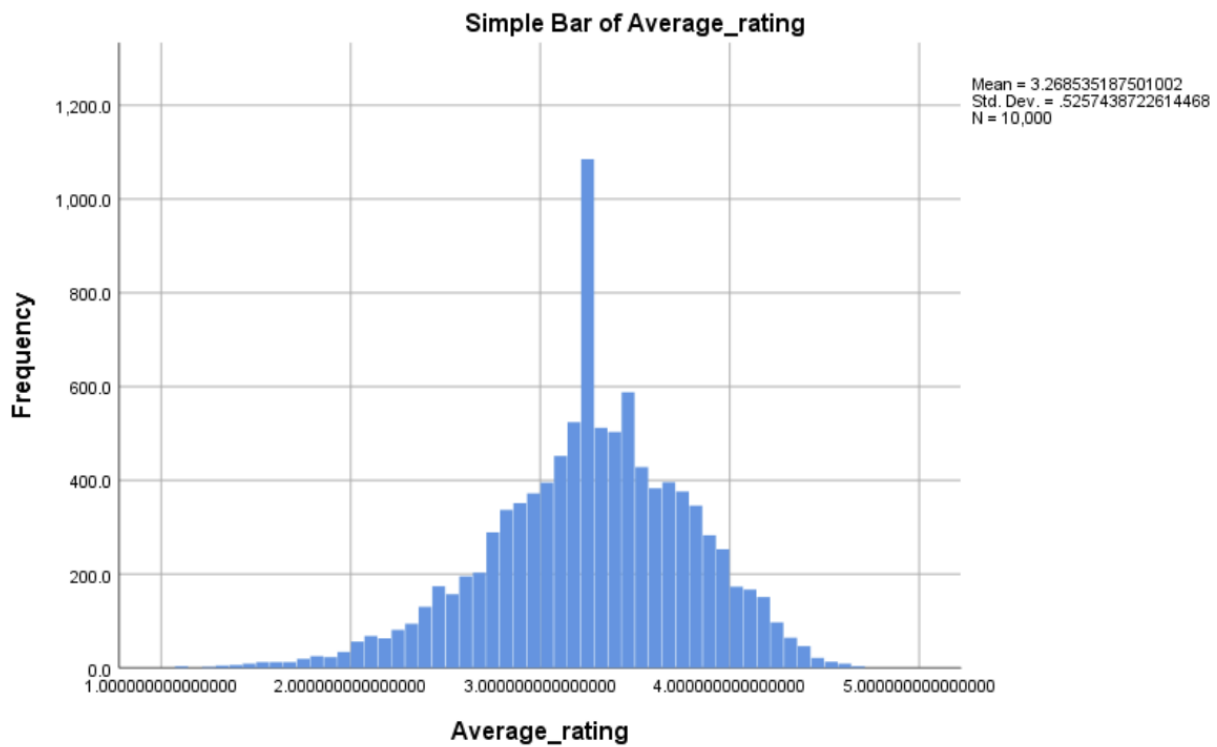


Figure 4: Average Rating

(Source: Imported from SPSS)

The histogram representation of the data clearly indicates that many of the greetings exist between three and four, with a sharp peak in the earliest part of the section between 3 and 4. Further representation of the rating, the possible reasons for the outliers are observed through the box plot analysis, as shown below.

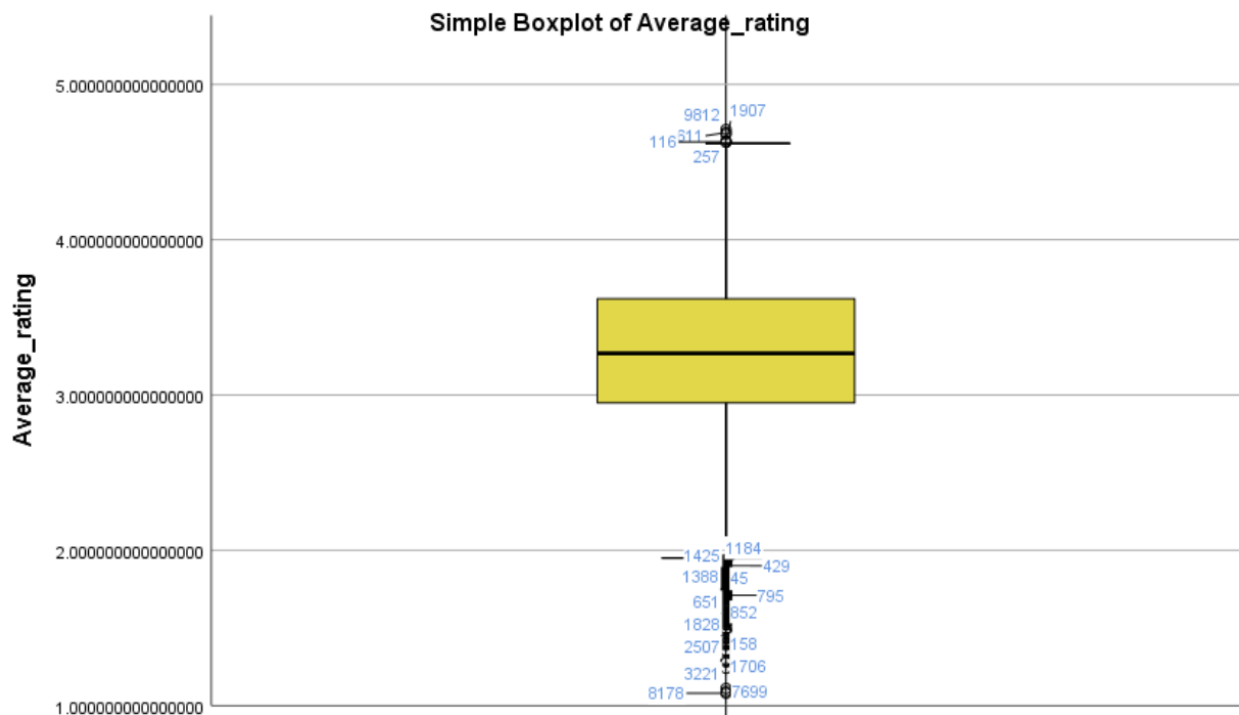


Figure 5: Boxplot representation of the rating

(Source: Imported from SPSS)

The box plot representation is also reflective that most of the observations lie between three and 3.8. There were a few observations that were observed below 2.0 and above 4.5. These are some of the partly recorded information which has a negligible influence on the data outcome.

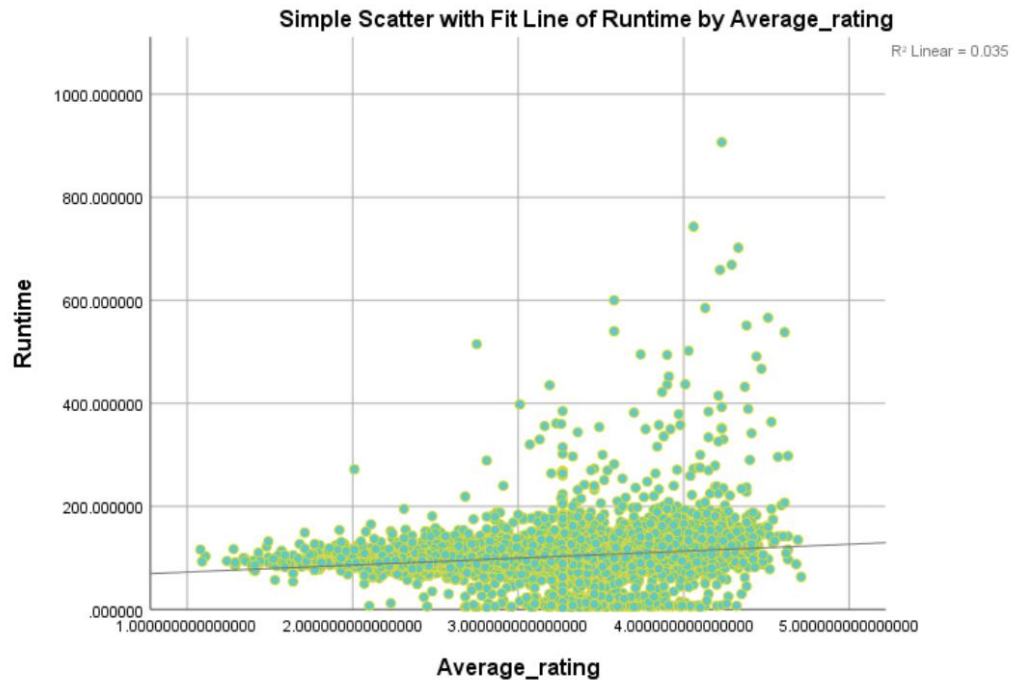


Figure 6: Relationship between runtime and average rating

(Source: Imported from SPSS)

It is observed that the higher the runtime, the greater the average rating provided, reflecting higher engagement over a significant runtime period. The result reflects the understanding from the customer engagement theory, which reflects the importance of generating higher value through the influence over the cognitive, emotional, and behavioural factors of the consumers. It signifies that achieving a higher rating could successfully keep the consumers engaged, and they feel value for the time they had invested in watching the film (Ng, Sweeney and Plewa, 2020). From the business perspective, the movies that are selected and are being shown on the platform have good content quality, which needs to be maintained in the future.

5. Statistical analysis in SPSS

5.1. Hypothesis testing through an independent t-test

The first part of the analysis includes a hypothesis testing conducted through an independent t-test.

H0 (null): The difference in the average rating between English and non-English films is not significant.

H1(Alternative): The difference in the average rating between English and non-English fans is significant.

Independent Samples Test										
		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
Average_rating	Equal variances assumed	27.710	.000	-23.009	9998	.000	-.298965724	.0129932830	-.324435174	-.273496274
	Equal variances not assumed			-24.878	3219.603	.000	-.298965724	.0120173056	-.322528068	-.275403380

Figure 7: Independent t-test result

(Source: Imported from SPSS results)

The homogeneity test is understood through the F value, which is 27.71, and it suggests that the homogeneity assumption is violated. In this situation, likely, the equality of the variance will not exist. Therefore, the t-statistic result for the equality of variance not assumed is considered, where the p-value is found to be less than 0.05, and the t-value is -24.878, rejecting the null hypothesis. It indicates that the mean average rating of the English movies is lower than the mean rating of the non-English movies. This is likely to occur because more than 80% of the data set includes English movies. The presence of diversity lowers the average value, which is reflected in this case (Röseler *et al.*, 2024). Hence, from the business perspective, the language remains an important factor for consumers to provide a critical perception of the movie.

5.2. Correlation Analysis

		Correlations		
		Runtime	Watches	Likes
Runtime	Pearson Correlation	1	.146**	.129**
	Sig. (2-tailed)		.000	.000
	N	10000	10000	10000
Watches	Pearson Correlation	.146**	1	.951**
	Sig. (2-tailed)	.000		.000
	N	10000	10000	10000
Likes	Pearson Correlation	.129**	.951**	1
	Sig. (2-tailed)	.000	.000	
	N	10000	10000	10000

** . Correlation is significant at the 0.01 level (2-tailed).

Figure 8: Correlation Results

(Source: Imported from SPSS)

The relationship highlighted that the most significant and strongest relationship is seen between watches and likes. The higher viewership is found to be related to the audience acceptance capacity. This is understood through the theory of social proof as consumers rely on social vision to generalise their perception and their choices that structure the market (Huang *et al.*, 2025). The other factors, like the relationship runtime, watches and likes, do not hold a strong relationship, even though there remains a significant relationship.

5.3. Movie description relationship with rating and language influence over engagement

Model Summary									
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	Change Statistics			Sig. F Change
						F Change	df1	df2	
1	.023 ^a	.001	.000	.5256296600	.001	5.346	1	9998	.021
a. Predictors: (Constant), Description_Length									

ANOVA ^a						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.477	1	1.477	5.346	.021 ^b
	Residual	2762.313	9998	.276		
	Total	2763.790	9999			
a. Dependent Variable: Average_rating						
b. Predictors: (Constant), Description_Length						

Figure 9: Regression Results

(Source: Imported from SPSS)

The regression analysis helped in determining the strength of the relationship between the description length and the average movie rating. The R-squared value of 0.001 indicates that only 0.1% change in the average rating is caused by the description length, reflecting the importance of providing a concise description (Wang, 2024). Further, the p-value is formed to be 0.021, which is less than 0.05, indicating that the relationship is significant. However, the model needs to consider other important variables, like the genre or language, to determine the audience perspective related to a movie.

Model Summary									
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.188 ^a	.035	.035	450384.367	.035	183.543	2	9997	.000

a. Predictors: (Constant), Runtime, language_num

ANOVA ^a					
Model		Sum of Squares	df	Mean Square	Sig.
1	Regression	7.446E+13	2	3.723E+13	183.543
	Residual	2.028E+15	9997	2.028E+11	.000 ^b
	Total	2.102E+15	9999		

a. Dependent Variable: Watches
b. Predictors: (Constant), Runtime, language_num

Figure 10: Influence of language on engagement

(Source: Imported from SPSS)

Identical understanding, highlighting that the runtime and the language have a moderate influence over the watching rate. The statistical significance is observed from the p-value being lower than 0.05. The moderate influence is observed through the R-squared value of 0.35, which suggests 35% of the watching rate is influenced by the runtime and the language of the movie. Therefore, the business or the content platform must focus more on providing attractive leadership by removing the language barrier in the movie and help widen the movie circle by incorporating a diversified group.

6. Insights and Recommendations

The observation from the analysis is highly influential in understanding the behaviour of the audiences. The application of statistical analysis, such as the t-test, reflects the influence of the language perspective on the average rating. The language is found to be a major factor influencing consumer behaviour, even though the result is highly biased towards English-language movies. The outcome of the analysis recommends that the business focus more on conducting market segmentation to determine the variation in the consumer perspective in terms of their preferences over language. The regression result indicated that the runtime and the language differences can influence the rate moderately. However, the audience engagement is dependent on the language,

run time, and the watch rate, reflecting the popularity factor in choosing the film. The best possible recommendation for the business to undertake in this situation is to implement a region-specific target market, which can be performed through market segmentation. It can engage a wide consumer segment into their target base by providing proper subtitles to promote good contains into a diversified market.

The outcome of this analysis is largely limited because a large portion of the observation has been provided in a textual format. Therefore, conducting quantitative analysis has remained largely limited to a few variables. It would be ideal for the future to conduct text analysis for generating an appropriate qualitative understanding to improve strategic decision-making.

References

- Alexakis, T., Adamopoulou, E., Peppes, N., Daskalakis, E. and Ntouskas, G. (2025). Evaluating Data Quality: Comparative Insights on Standards, Methodologies, and Modern Software Tools. *Electronics*, 14(15), p.3038.
- Huang, W., Wang, X., Zhang, Q., Han, J. and Zhang, R. (2025) 'Beyond likes and comments: How social proof influences consumer impulse buying on short-form video platforms', *Journal of Retailing and Consumer Services*, 84, p.104199.
- Mu, W., Cardelli, R. and Ferrari, S. (2026). Data Preprocessing Techniques for Machine Learning Towards Improving Building Energy Performance: A Systematic Review. *Energies*, 19(6), p.1561.
- Ng, S.C., Sweeney, J.C. and Plewa, C. (2020) 'Customer engagement: A systematic review and future research priorities', *Australasian Marketing Journal*, 28(4), pp.235-252.
- Röseler, L., Felser, G., Asberger, J. and Schütz, A. (2024) 'The Effect of Variety on Perceived Quantity: Failures to Replicate Redden and Hoch (2009)', *Meta-Psychology*, 8.
- Wang, J.Q. (2024) 'Investigating the Effectiveness of Short Form Media Advertisements Compared to Long Form Media Advertisements', *Advances in Economics, Management and Political Sciences*, 89(1), pp. 187–192.