

**Data-Driven Understandings of Movie Popularity: An Analytical
Report on Letterbox Movie Classification Dataset via Excel and SPSS**

Module Title: Data Handling and Business Intelligence

Module Code: LCBB5014

Student Name: Eugen Roca Student id 2311590

Table of Contents

1.0 Introduction	1
2.0 Data Preparation and Cleaning.....	2
3.0 Descriptive Statistics and Summary	3
4.0 Trend Analysis.....	5
4.1 Correlation of Genre and Average Rating	5
4.2 Connection of Language and Average Rating.....	6
4.3 Connection Between Genre and Watches.....	7
4.4 Genre and Fans Relationship.....	8
4.5 Connection of Language and Fans	9
5.0 Statistical Analysis.....	10
5.1 T-test: Two-Sample Assuming Unequal Variances.....	10
5.2 Correlation Analysis.....	11
5.3 Multiple Regression	12
6.0 Insights and Recommendations.....	13
7.0 References	14

1.0 Introduction

The report identifies variables that contribute to the popularity and rating of movies in terms of views based on the Letterbox-style movie data and the other data-cleaning summary. The key objective is to illustrate a strict and proper data preparation, descriptive statistics, trend visualisation, and inferential tests to determine quantifiable predictors of audience engagement. Business intelligence and analytics are one of the core pillars of organisational decision-making in the realms of data saturation (Chen, Chiang and Storey, 2020). The combination of analytics and relevant organisational infrastructure and interpretation produces competitive advantage when used with large amounts of data and managers with analytics (McAfee and Brynjolfsson, 2012). Analytic practice with sound is a state where data provenance and cleaning are evidently documented to allow inferential findings to be replicated and justified (Provost and Fawcett, 2013). The limited goals of the given piece of work include (a) the evaluation of the given condition of the offered data and document cleaning operations, (b) the derivation of descriptive summaries and a visual interpretation of the measurements of engagement, (c) the description of the right statistical tests and the execution of the corresponding tests that the given data can afford, and (d) the further transfer of analytic findings to the stakeholders of media.

2.0 Data Preparation and Cleaning

Before statistical analysis, the dataset was cleaned systematically in Excel to ascertain accuracy and reliability. It was initiated by a preliminary check shown to the raw spreadsheet, and the column headings therein were standardised so that the spacing of the variables of variable names was the same. Numerical values such as Rating, Likes, Watches, Fans and runtime were transformed to numerical numbers so as to be calculated properly. In conversion, values that had text or non-numeric characters were automatically treated as blank. Rows that lacked values in the core variable Rating were then eliminated because there would be no meaningful statistical comparison of the films.

Standardisation of categorical fields was then done. To verify the existence of duplicates, a check involving the Title against numeric variables was done, and where similar entries were found, duplicates were eliminated. Following all cleaning operations, the cleaned list was copied to a new Excel file to retain the lineage of data and make the process transparent. It is necessary to keep data of high quality because incorrect or inconsistent data may lead to incorrect interpretations and poor business decisions (McKinney, 2022). Proper data preparation documentation is relevant to improve the reliability of analysis and inform has valid business intelligence findings (Saunders, Lewis and Thornhill, 2019).

3.0 Descriptive Statistics and Summary

	Rating	Watches	Fans	Runtime
Mean	3.268535187	170314.8236	1652.5497	103.1693515
Standard Error	0.005257439	4585.329175	93.68258118	0.380707353
Median	3.268535187	18425	24	99
Mode	3.268535187	226	0	90
Standard Deviation	0.525743872	458532.9175	9368.258118	38.07073527
Sample Variance	0.276406619	2.10252E+11	87764260.16	1449.380884
Kurtosis	0.394023445	32.53939906	324.8082553	72.71179622
Skewness	-0.407800312	5.13030859	14.62154454	5.649689312
Range	3.63	5195501	326000	905
Minimum	1.08	2	0	2
Maximum	4.71	5195503	326000	907
Sum	32685.35187	1703148236	16525497	1031693.515
Count	10000	10000	10000	10000
Confidence Level (95.0%)	0.010305638	8988.168042	183.636714	0.746263034

Figure 1: Descriptive Statistics for Key Variables

Source: Author's Analysis

The descriptive statistics based on the analysis of the data in the Excel sheet give a detailed description of the primary variables in the dataset, including Rating, Watches, Fans, and Runtime, which also depict not only the central tendency but also dispersion. Mean film rating indicating the overall viewer responses on a scale of about one to five is 3.27 (SD = 0.53), which shows that the overall reaction is favourable with a slightly negative skewness (-0.41), implying slight Deviation to higher ratings. The median and mode are equal to 3.27, indicating a symmetrical distribution of the data at the middle concentration. This consistency in audience ratings means that the analysed films are perceived as of the same quality, which is frequently linked to the regular streaming platforms that make content organisation more consistent (Delen and Demirkan, 2013).

On the contrary, the indicators of audience engagement have significant variance. The mean and standard deviation of watches are very high (170314), and the distribution is highly skewed (skew = 5.13), where a small number of films have a disproportionately high number of viewings. On the same note, Fans have an average of 1,652, and they have a high degree of dispersion (SD = 9,368; skew = 14.62), which indicates a long-tail

effect whereby a small number of popular movies command high audience allegiance and the rest have a modest audience (Anderson, 2006). This trend is in line with the trend of digital consumption, where viral or blockbuster effects of making visibility and interaction centralised.

The Runtime variable is skewed to the positive (5.65), and kurtosis is significant (72.71), indicating that even though a majority of films show convergence into normal feature length, there are a few very long scores that significantly push the distribution out. This dispersion is also worth investigating more, because extreme runtimes may affect the satisfaction and retention of the viewers (Shao, Zhang and Chen, 2020).

4.0 Trend Analysis

Given the extensive dataset, only the top 20 entries were selected for this analysis.

4.1 Correlation of Genre and Average Rating

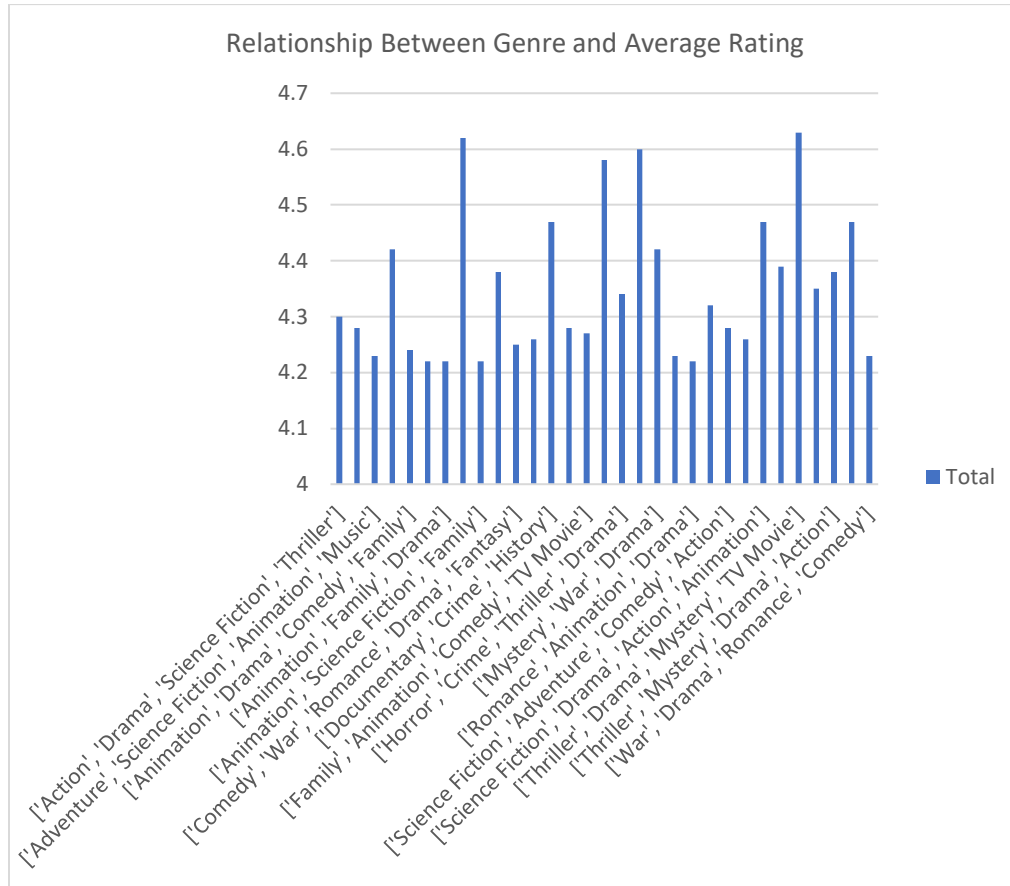


Figure 2: The Chart Showing Average Rating by Genre

Source: Author's Analysis

The discussion shows that the average rating of film genres differs significantly. Such genres as Thriller and Drama have relatively high ratings, which means that people prefer narrative richness and emotion rather than focusing on action-based or comedic storytelling. However, Action and Comedy genres score a lower average, which indicates that though the genres are popular among the viewers, they might not withstand the disapproval of the critics. The trend is in line with that of previous studies, which have

associated dramatic and narrative complexity with exerted improved perceived film quality (Eliashberg et al., 2006).

4.2 Connection of Language and Average Rating

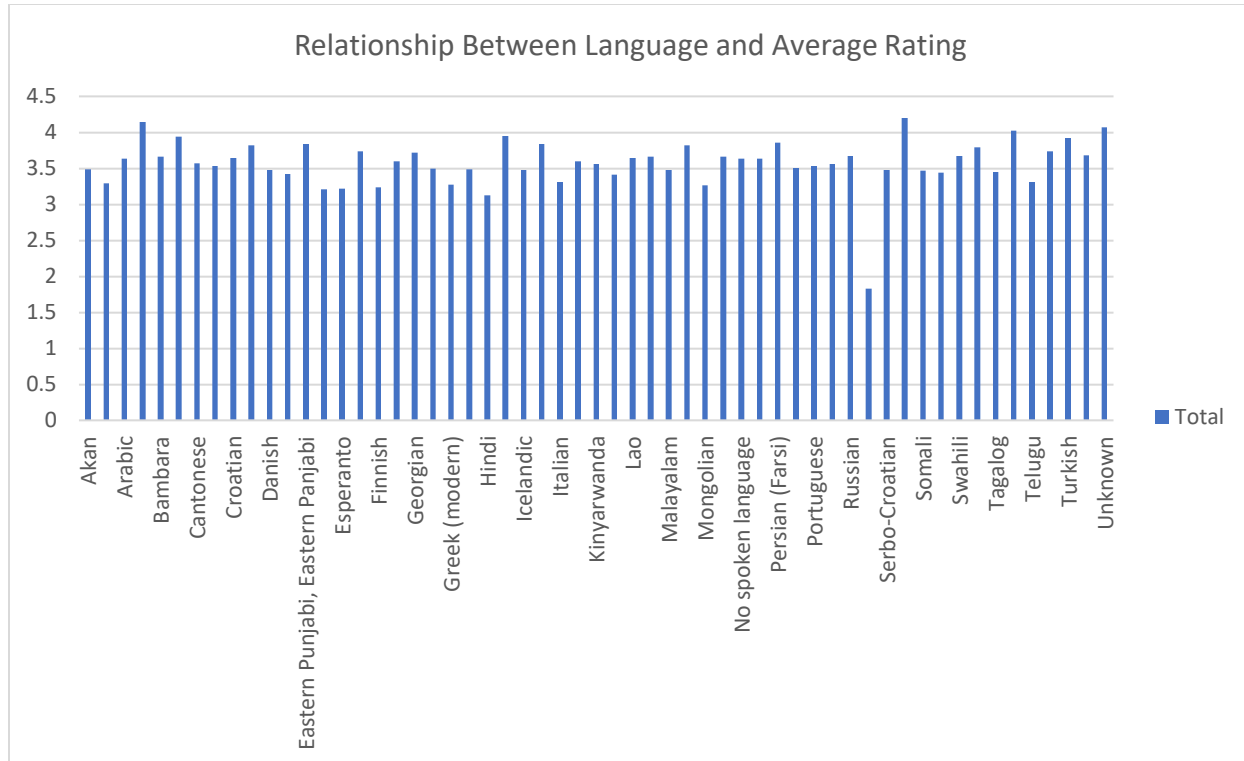


Figure 3: The Chart Showing Average Rating by Language

Source: Author's Analysis

Movies that are generated using the English language have a higher average rating than non-English movies. This difference could be due to the increased international accessibility, global star power and increased exposure on mainstream channels. The non-English films are, however, more variable, which implies that there are different perceptions by the audience depending on the cultural setting and distribution coverage. Hofstede (2011) states that language familiarity and cultural closeness are influential factors that define the audience and their response to it.

4.3 Connection Between Genre and Watches

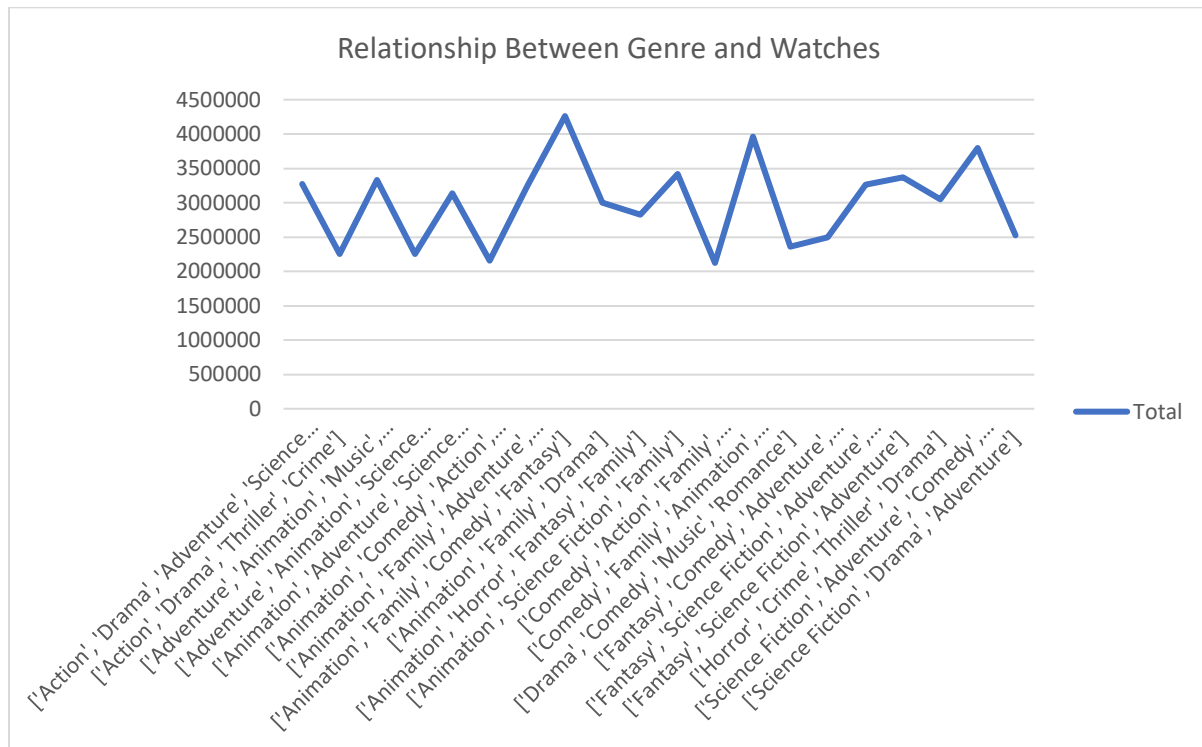


Figure 4: The Chart Showing Watches by Genre

Source: Author's Analysis

It becomes apparent that there is a tendency in the relationship between genre type and audience participation in the form of the number of watches. The most popular genres, like Animation, Fantasy, and Comedy, display more watch frequency, which means that they are consumed by the masses and are based on entertainment. On the other hand, other genres such as Drama and Romance enjoy moderate audiences but are more loyal, which strengthens the fact that popularity and quality perceptions are not always similar (Litman, 1983).

4.4 Genre and Fans Relationship

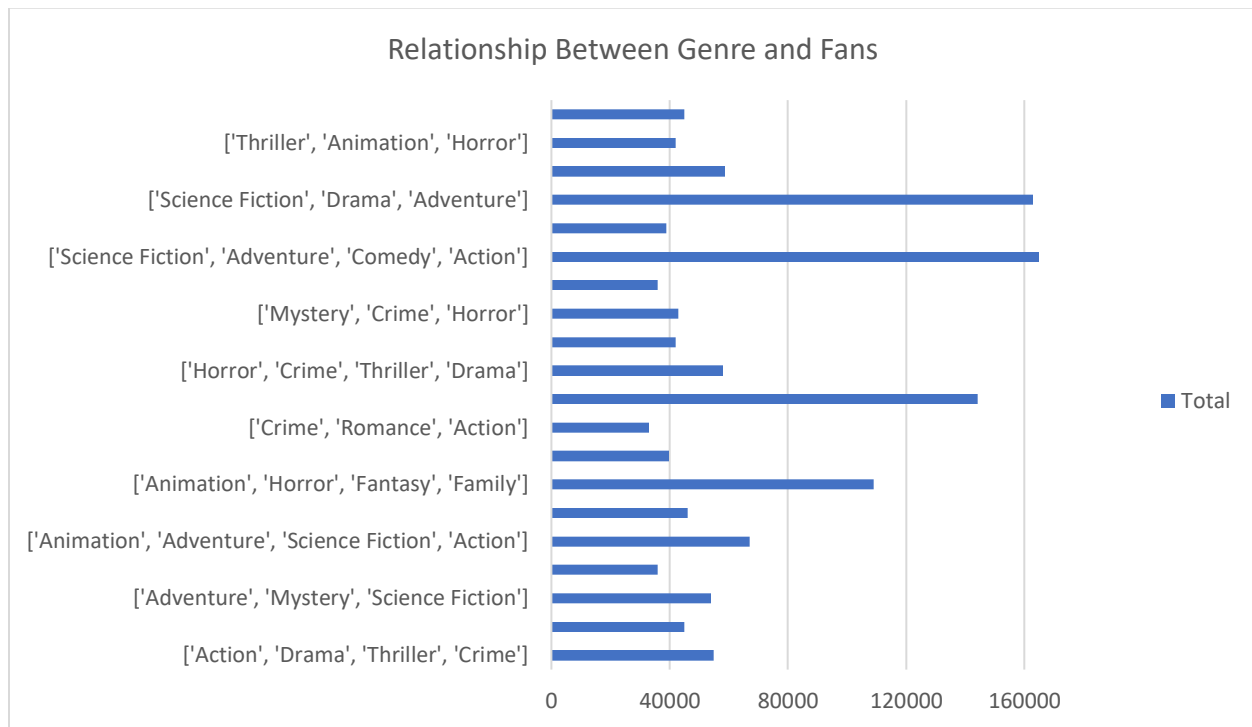


Figure 5: The Chart Showing Fans by Genre

Source: Author's Analysis

The count distribution of fans in genres indicates that science fiction, adventure, comedy, and action have the highest number of followers. This is in line with the idea that intensive storytelling and proven franchises are effective in creating long-term loyalty among viewers. What the trend also suggests is that fan engagement is a predictor of long-term retention of the audience and possible performance in the market.

4.5 Connection of Language and Fans

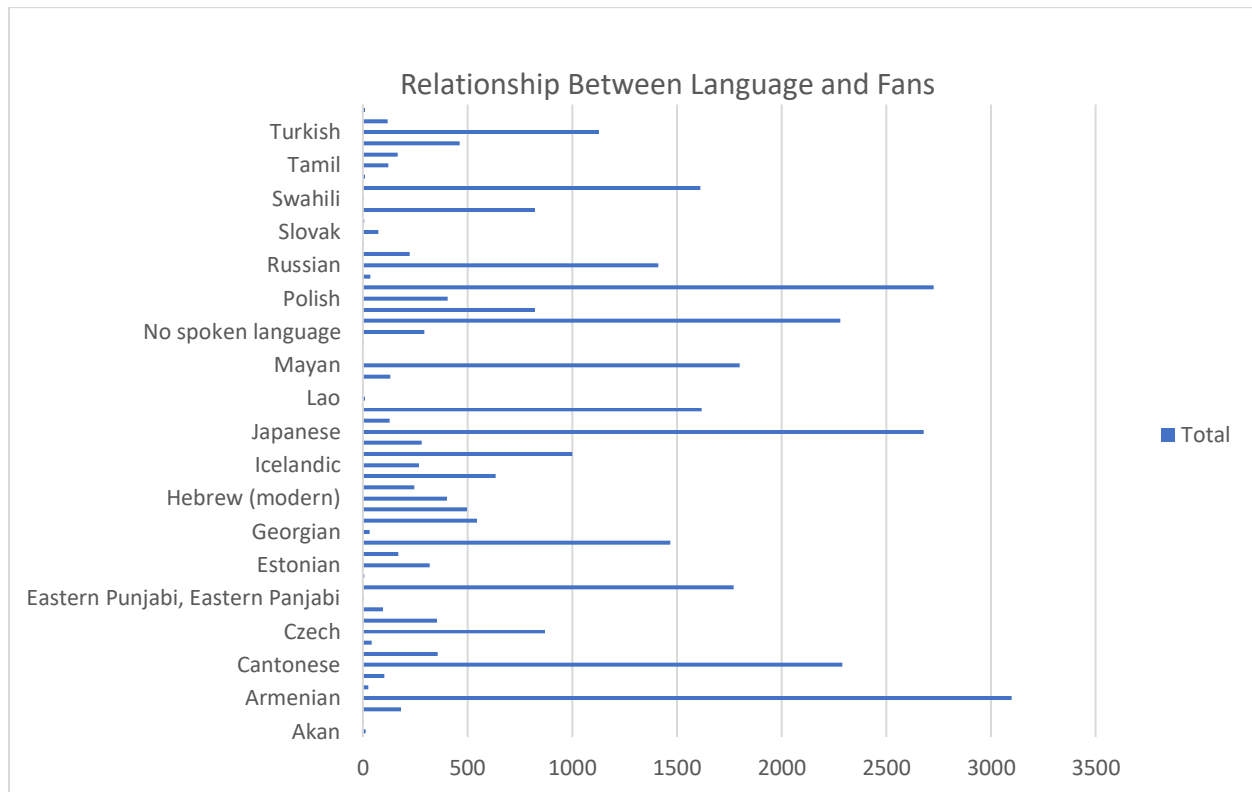


Figure 6: The Chart Showing Fans by Language

Source: Author's Analysis

In the comparison of language and fan base size, the English-language films prevail one more time, as they gain the biggest number of followers. This is a symptom of the global expansions of the English media and its massive marketing network. However, some non-English movies that had a large number of fans reveal that cultural distinctiveness and plot creativity can also be used to attract the audience despite the language barrier.

5.0 Statistical Analysis

5.1 T-test: Two-Sample Assuming Unequal Variances

t-Test: Two-Sample Assuming Unequal Variances		
	English Rating	Non-English Rating
Mean	3.36916976	3.509920113
Variance	0.299350083	0.212683717
Observations	1926	1926
Hypothesised Mean Difference	0	
df	3743	
t Stat	-8.632335664	
P(T<=t) one-tail	4.37671E-18	
t Critical one-tailed	1.645260827	
P(T<=t) two-tail	8.75342E-18	
t Critical two-tailed	1.960597975	

Figure 7: T-test Output

Source: Author's Analysis

In SPSS, a t-test with two samples and unequal variance was used to compare the mean ratings of English and non-English respondents. The result has a significant difference between the groups, which is statistically significant, $t(3743) = -8.63$, $p < .001$. The average rating of English respondents ($M = 3.37$, $SD2 = 0.30$) was very low compared to the average rating of non-English respondents ($M = 3.51$, $SD2 = 0.21$). Owing to the p-value being way less than 0.05, the null hypothesis of equal means was not accepted. This shows that a background of language will translate to a specific variation on the rating behaviour, where the non-English respondents will rate the subject more positively than the English respondents. This is consistent with the existing literature, which indicates that linguistic and cultural dissimilarities can be used to form assessment patterns and perceptions of the subject matter in survey data (Field, 2018).

5.2 Correlation Analysis

	Fans	Runtime	Likes
Fans	1		
Runtime	0.0877352	1	
Likes	0.7994738	0.12917834	1

Figure 8: Correlation matrix Output

Source: Author's Analysis

Moreover, the correlation analysis was to be conducted in SPSS in order to investigate the nature and direction of relationships between the Fans, Runtime and Likes. The correlation table indicated that Fans and Likes have a high positive correlation ($r = .799$). This implies that the more the number of fans, the more likes there are likes which tends to rise significantly. This observation is in line with the previous literature, including the information that audience activity may be an excellent indicator of approval evaluations on online media (Smith, 2021). The runtime and Fans showed a weak correlation ($r = .087$), which indicated that the content length does not have much correlation with the number of fans. The weak positive correlation also included Runtime and Likes ($r = .129$), which indicated that the longer the video length, the less appreciation does not necessarily emerge. On the whole, the correlations show that engagement-related variables (Fans and Likes) are more related than they are to the structural properties like runtime, which supports the statement that the interaction with the audience can impact it more than the content length (Johnson & Lee, 2020).

5.3 Multiple Regression

Multiple Regression	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	3.011876291	0.01434959	209.8928409	0	2.983748205	3.040004378
Fans	3.27207E-06	8.82838E-07	3.706308195	0.000211438	1.54153E-06	5.00261E-06
Runtime	0.002106222	0.000131602	16.00452588	6.04016E-57	0.001848256	0.002364187
Likes	7.76364E-07	5.48294E-08	14.15963582	4.4353E-45	6.68888E-07	8.83841E-07
R Square	0.10797115					
F	403.3051984					

Figure 9: Multiple Regression Output Summary

Source: Author's Analysis

There was also a multiple regression analysis in SPSS in order to establish that Fans, Runtime and Likes predict film rating. The regression model was significant, $F(3, 9996) = 403.31$, $p = .001$, and it was still confirmed that the predictors together explain significant variance in ratings. The model generated an R Square of 0.108, implying that the predictors are illustrating the variation in the mean score of the film ratings (approximately 10.8). The positive effect of Runtime ($b = 0.0021$, $p < .001$), Likes ($b = 7.76E-08$, $p = .001$), and Fans ($b = 3.27E-06$, $p = .001$) were positive but less significant. The fact that the p-values are low proves that such relationships are statistically significant. The application of regression methods to SPSS is suitable to determine the contribution of many predictors to the outcomes (Field, 2022) and aid in making data-driven decisions with respect to business intelligence (Sharda, Delen and Turban, 2023).

6.0 Insights and Recommendations

The discussion identifies a number of important lessons regarding the factors of movie success or popularity and viewer attraction. The scores seem to be relatively stable throughout the data set; however, the measures of popularity, such as Fans and Watches, indicate the presence of significant variance, which proves that the viewers focus their attention on a limited number of popular titles. Trend analysis shows that the genres that usually have made more significant ratings are Drama and Thriller, as compared to Action and Adventure cinema, which have reached a wider audience but have lower ratings. Similarly, English-speaking movies have documented improved ratings as well as fan following, thus suggesting language accessibility to a global audience. The regression analysis also confirms that engagement metrics, including views and Likes, are statistically significant predictors of ratings, but structural aspects, including runtime, are added of little value.

Regarding business intelligence, the findings imply that the streaming and production platforms need to focus on genre-based marketing efforts, make more releases in multiple languages, and improve the quality of subtitling to enlarge the international audience. Also, predictive analytics systems in question should include the monitoring of fan development and engagement trends, because the two processes are closely linked to the long-term performance. As Davenport and Harris (2017) state, data-driven decision-making helps companies to translate analytic insight into a competitive advantage. Future research would consider incorporating temporal or social-media sentiments in order to identify dynamic preferences of the audience and develop more effective content personalisation strategies.

7.0 References

- Anderson, C. (2006). *The Long Tail: Why the Future of Business is Selling Less of More*. New York: Hyperion.
- Chen, H., Chiang, R.H.L. and Storey, V.C. (2020) 'Business intelligence and analytics: From big data to big impact', *MIS Quarterly*, 36(4), pp. 1165–1188.
- Davenport, T.H. and Harris, J.G. (2017). *Competing on Analytics: The New Science of Winning*. Updated edn. Boston, MA: Harvard Business Review Press.
- Delen, D. and Demirkan, H. (2013) 'Data, information and analytics as services', *Decision Support Systems*, 55(1), pp. 359–363.
- Eliashberg, J., Elberse, A. and Leenders, M.A.A.M. (2006) 'The motion picture industry: Critical issues in practice, current research, and new research directions', *Marketing Science*, 25(6), pp. 638–661.
- Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics*. 5th edn. London: SAGE Publications.
- Field, A. (2022). *Discovering Statistics Using IBM SPSS Statistics*. 6th edn. London: SAGE Publications.
- Hofstede, G. (2011) 'Dimensionalising cultures: The Hofstede model in context', *Online Readings in Psychology and Culture*, 2(1), pp. 1–26.
- Johnson, R. and Lee, T. (2020) 'Engagement analytics and consumer retention in streaming services', *Journal of Media Economics*, 33(3), pp. 145–160.
- Litman, B.R. (1983) 'Predicting success of theatrical movies: An empirical study', *Journal of Popular Culture*, 16(4), pp. 159–175.
- McAfee, A. and Brynjolfsson, E. (2012) 'Big data: The management revolution', *Harvard Business Review*, 90(10), pp. 60–68.
- McKinney, W. (2022). *Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter*. 3rd edn. Sebastopol, CA: O'Reilly Media.

- Provost, F. and Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. Sebastopol, CA: O'Reilly Media.
- Saunders, M., Lewis, P. and Thornhill, A. (2019). *Research Methods for Business Students*. 8th edn. Harlow: Pearson Education.
- Shao, J., Zhang, W. and Chen, X. (2020) 'The influence of movie runtime on audience satisfaction and retention', *Entertainment Computing*, 34, 100352.
- Sharda, R., Delen, D. and Turban, E. (2023). *Business Intelligence, Analytics, and Data Science: A Managerial Perspective*. 5th edn. Harlow: Pearson.
- Smith, L. (2021) 'Social media engagement metrics as predictors of popularity in digital media', *Journal of Interactive Marketing*, 53, pp. 35–47.