



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

LCBB5014

Data Handling & Business Intelligence

Movie Classification

Student Name: Cristian Valenzuela Garcia

Student ID: 2312699

Lecturer name: Dr Said Usman

Table of Contents

1.Introduction	2
2. Data Cleaning and Preparation	2
3. Descriptive Statistics and Summary.....	4
3.1 Descriptive Statistics of Fans	5
3.2 Descriptive Statistics of Watches	5
3.3 Descriptive Statistics of Runtime	6
3.4 Descriptive Statistics of Average Rating.....	7
4. Trend Analysis and Visualization	7
4.1 Trends correlated with language	8
4.2 Genres associated with high average ratings.....	9
5. Statistical Analysis using SPSS.....	10
5.1 Difference between English and non-English films by Perform t-test.	11
5.2 Correlations between variables with longer runtimes determine tend to have more watches or likes.....	13
5.3 Regression Analysis	16
6. Recommendations	17
7. Conclusion	17
8. References	18

1.Introduction

The purpose of this Assessment is to apply the skills acquired in analysing and developing a database, ensuring that the information is presented in a clear and accessible manner. Trinity Saint David University of Wales provides a real dataset concerning the classification of movies, which includes variables such as title, genre, director, language, and ranking. This dataset serves as the foundation for exploring and analysing information, identifying trends, examining viewer behaviour patterns, and investigating the factors that contribute to the popularity of films.

The process will initially be conducted in Excel, where the data will be cleaned and organized. This stage involves adding missing values, removing irrelevant information, and ensuring consistency across the dataset. Once the data has been refined, descriptive statistics will be generated, including averages of movie views, fan counts, and rankings. These measures will provide a clearer understanding of the dataset and highlight key insights into audience preferences.

Subsequently, the analysis will be replicated in SPSS to validate results and apply more advanced statistical techniques. To enhance interpretation, dashboards will be created, incorporating bar charts, line graphs, and other visualizations. These graphical representations will facilitate the communication of findings, making complex information easier to understand and supporting evidence-based conclusions.

2. Data Cleaning and Preparation

Data cleaning is the process of detecting, correcting, or deleting errors and adding missing values to ensure data is accurate, consistent, and useful. Its objective is enabling statistical analysis and business decisions, improving efficiency and reflecting reality, Tableau (2024).

In this case, Excel shows film database such as movie titles, Director, average, genres, views, language, and ranking fragmented.is concentrated in 10001 rows and columns from A to Q. database appear problems of codification characters as Lowestâ, Mediumâ, Highestâ, due to the fact the file is in a format with a different encoding. For example, in CSV. In this case Excel do not identify the codification and interpret in bad symbols. It does not affect the information, it's just visual. As consequence it contributes to enhancing the clarity and usefulness of both the message and the data.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1		Film_title	Director	Average_rating	Genres	Runtime	Original_language	Studios	Watches	List_appears	Likes	Fans	Lowest	Medium	Highest	Total_ratings			
2	0	The Fan	Eckhart Schmidt	3.57	Horror, Drama	92	German	The hero-woman Barbara Moore Workshop	12841	11032	3763	79	129	1660	714	9042			
3	1	Mad Max: Fury Road	George Miller	4.18	Adventure, Science Fiction	121	English	An apocalypse Warner Bros. Pictures, Village Roadshow Pictures	2488721	331681	927538	38000	12530	158356	511140	1682389			
4	2	Suspiria	Dario Argento	3.93	Horror	99	English	An American Seda Spettacoli	618171	220094	219548	13000	2814	53427	88628	443757			
5	3	Lost in Translation	Sofia Coppola	3.79	Drama, Comedy, Fantasy	102	English	Two lost souls American Zoetrope, Element	1596190	254180	493248	38000	15167	155110	193717	1076949			
6	4	Akira	Katsuhiro Oton	4.28	Animation, Action	124	Japanese	A secret mission MBS, Sumitomo Corporation	873068	175586	341567	21000	1822	40850	196532	600721			
7	5	Things to Come	Mia Hansen-Løve	3.77	Drama	102	French	Nathalie te ARTE France Cinéma, De	34777	14370	8585	192	90	4274	1677	24383			
8	6	Big Trouble in Little China	John Carpenter	3.72	Comedy, Adventure	100	English	Truck driver TAFT Entertainment Pictures	283873	65558	85788	3800	1250	30582	24938	183927			
9	7	Star Wars	George Lucas	4.17	Adventure, Action	121	English	Princess Leia Lucasfilm Ltd., 20th Century	3014320	391415	1057816	30000	6041	161969	550396	1780203			
10	8	The Third Man	Carol Reed	4.25	Mystery, Thriller	105	English	In postwar London Films Productions	228162	101285	70932	3400	351	11549	39520	152601			
11	9	The Big Sick	Michael Showa	3.76	Drama, Comedy, Fantasy	120	English	Pakistan-born FilmNation Entertainment,	382201	58859	93489	504	1410	36987	18997	238309			
12	10	Jodorowsky's Frank Pavich	Frank Pavich	3.96	Documentary	90	English	Shot in France Snowfort Pictures, Cam	85388	16745	25616	188	209	6825	6696	56507			
13	11	Get Out	Jordan Peele	4.15	Horror, Mystery, Thriller	104	English	Chris and Jordan Monkeypaw Productions, B	4376009	454087	1867622	33000	10251	272655	943393	3121586			
14	12	Lawrence of Arabia	David Lean	4.38	Adventure, War, History	228	English	The story of Lord Horizon Pictures	369742	151348	116793	10000	1080	16158	86313	242207			
15	13	Free Fire	Ben Wheatley	3.16	Action, Crime, Mystery	90	English	Set in Boston Rook Films, Protagonist Pictures	109849	29433	21524	83	1528	22924	1853	81191			
16	14	Sing Street	John Carney	4.04	Romance, Drama	105	English	A boy growing up Likely Story, FilmWave, Disney	390910	90763	145995	10000	846	27808	55425	273809			
17	15	Up	Pete Docter	4.1	Animation, Adventure	96	English	Carl Fredrick Pixar	3922213	263407	1136383	9100	6354	230485	591788	2091274			
18	16	Blade Runner	Ridley Scott	4.1	Thriller, Drama, Science Fiction	118	English	In the smoggy Shaw Brothers, The Ladd Company	1658845	301778	573522	51000	6694	111183	309621	1095317			
19	17	Annie Hall	Woody Allen	4.04	Drama, Romance	93	English	New York-born United Artists, Rollins-Joffe	454236	121871	119958	6400	2829	30173	60635	273663			
20	18	The Truman Show	Peter Weir	4.22	Comedy, Drama	103	English	Every second Paramount Pictures, Scott	4382819	361732	1777004	71000	5764	231650	930042	2926289			
21	19	Wonder Woman	Patty Jenkins	3.24	Action, Fantasy, Adventure	141	English	An Amazonian Atlas Entertainment, Cruel	1721440	147437	320485	1600	25306	239158	70337	946010			
22	20	Plan 9 from Outer Space	Edward D. Wood	2.24	Science Fiction, Horror	79	English	In California Reynolds Pictures	61992	20589	11196	121	6724	4949	2272	35588			
23	21	Whiplash	Damien Chazal	4.42	Drama, Music	107	English	Under the Bold Films, Blumhouse Productions	3925603	481379	1848733	154000	7390	146378	1272739	2963756			

According to Sultana, (2025). Transformation in Excel refers to modifying data structure to meet statistical assumptions or facilitate calculations. Methods range from simple tasks, such as changing text format or replacing numbers, to advanced techniques like applying logarithms, roots, or normalization, ensuring accurate analysis and reliable results.

Suggested by Siddiqui, (2024). data removal is a part of the process of data cleaning in excel, is based on delete data from worksheet, of which are incorrect, irrelevant, duplicates or incomplete. As a consequence, improve quality of data and guarantee that the analysis is dependable.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
		Film_title	Director	Average_rating	Genres	Runtime	Original_language	Description	Studios	Watches	List_appears	Likes	Fans	Lowest	Medium	Highest
	0	The Fan	Eckhart Schmidt	3.57	Horror, Drama	92	German	The hero-woman Barbara Moore Workshop		12841	11032	3763	79	129	1660	714
	1	Mad Max: Fury Road	George Miller	4.18	Adventure, Science Fiction	121	English	An apocalypse Warner Bros. Pictures, Village Roadshow Pictures		2488721	331681	927538	38000	12530	158356	511140
	2	Suspiria	Dario Argento	3.93	Horror	99	English	An American Seda Spettacoli		618171	220094	219548	13000	2814	53427	88628
	3	Lost in Translation	Sofia Coppola	3.79	Drama, Comedy, Fantasy	102	English	Two lost souls American Zoetrope, Element		1596190	254180	493248	38000	15167	155110	193717
	4	Akira	Katsuhiro Oton	4.28	Animation, Action	124	Japanese	A secret mission MBS, Sumitomo Corporation		873068	175586	341567	21000	1822	40850	196532
	5	Things to Come	Mia Hansen-Løve	3.77	Drama	102	French	Nathalie te ARTE France Cinéma, De		34777	14370	8585	192	90	4274	1677
	6	Big Trouble in Little China	John Carpenter	3.72	Comedy, Adventure	100	English	Truck driver TAFT Entertainment Pictures		283873	65558	85788	3800	1250	30582	24938
	7	Star Wars	George Lucas	4.17	Adventure, Action	121	English	Princess Leia Lucasfilm Ltd., 20th Century		3014320	391415	1057816	30000	6041	161969	550396
	8	The Third Man	Carol Reed	4.25	Mystery, Thriller	105	English	In postwar London Films Productions		228162	101285	70932	3400	351	11549	39520
	9	The Big Sick	Michael Showa	3.76	Drama, Comedy, Fantasy	120	English	Pakistan-born FilmNation Entertainment,		382201	58859	93489	504	1410	36987	18997
	10	Jodorowsky's Frank Pavich	Frank Pavich	3.96	Documentary	90	English	Shot in France Snowfort Pictures, Cam		85388	16745	25616	188	209	6825	6696
	11	Get Out	Jordan Peele	4.15	Horror, Mystery, Thriller	104	English	Chris and Jordan Monkeypaw Productions, B		4376009	454087	1867622	33000	10251	272655	943393
	12	Lawrence of Arabia	David Lean	4.38	Adventure, War, History	228	English	The story of Lord Horizon Pictures		369742	151348	116793	10000	1080	16158	86313
	13	Free Fire	Ben Wheatley	3.16	Action, Crime, Mystery	90	English	Set in Boston Rook Films, Protagonist Pictures		109849	29433	21524	83	1528	22924	1853
	14	Sing Street	John Carney	4.04	Romance, Drama	105	English	A boy growing up Likely Story, FilmWave, Disney		390910	90763	145995	10000	846	27808	55425
	15	Up	Pete Docter	4.1	Animation, Adventure	96	English	Carl Fredrick Pixar		3922213	263407	1136383	9100	6354	230485	591788
	16	Blade Runner	Ridley Scott	4.1	Thriller, Drama, Science Fiction	118	English	In the smoggy Shaw Brothers, The Ladd Company		1658845	301778	573522	51000	6694	111183	309621
	17	Annie Hall	Woody Allen	4.04	Drama, Romance	93	English	New York-born United Artists, Rollins-Joffe		454236	121871	119958	6400	2829	30173	60635
	18	The Truman Show	Peter Weir	4.22	Comedy, Drama	103	English	Every second Paramount Pictures, Scott		4382819	361732	1777004	71000	5764	231650	930042
	19	Wonder Woman	Patty Jenkins	3.24	Action, Fantasy, Adventure	141	English	An Amazonian Atlas Entertainment, Cruel		1721440	147437	320485	1600	25306	239158	70337
	20	Plan 9 from Outer Space	Edward D. Wood	2.24	Science Fiction, Horror	79	English	In California Reynolds Pictures		61992	20589	11196	121	6724	4949	2272
	21	Whiplash	Damien Chazal	4.42	Drama, Music	107	English	Under the Bold Films, Blumhouse Productions		3925603	481379	1848733	154000	7390	146378	1272739

In the previous screenshots it can observe a new format without strange symbols, moreover, it allows us to visualize through filters with arrow shape nest to variables, where to select the part, we are interested in analysing, for example, in English version or foreign version.

3. Descriptive Statistics and Summary

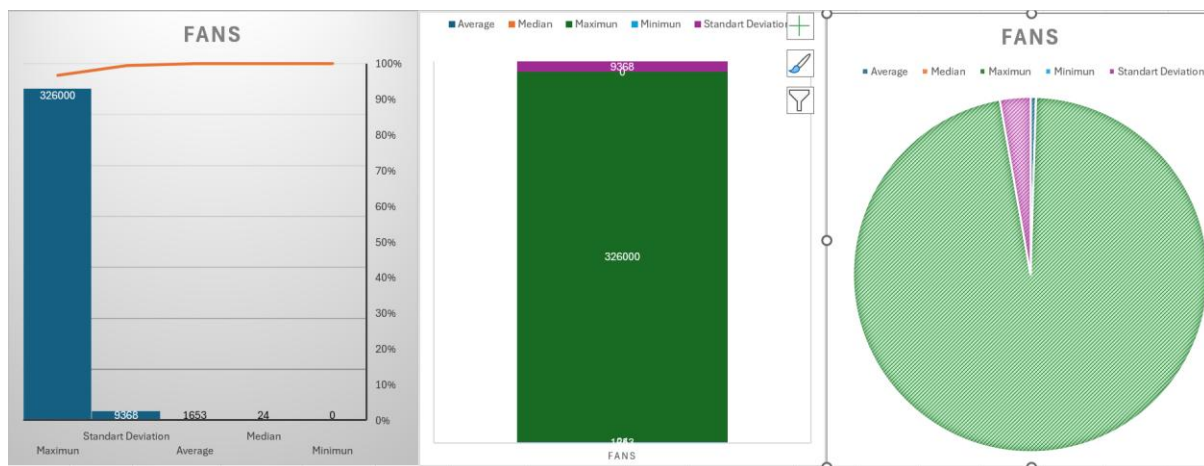
In this section of the report, Excel and SPSS are used to generate measures that summarize key variables, allowing the data to be understood without the need for inferences. According to Alabi and Bukola (2023), the purpose of descriptive statistics is to condense large volumes of information into representative numerical and graphical values that facilitate interpretation. The objectives of descriptive statistics include organizing extensive datasets through tables, summarizing trends and measures of dispersion to synthesize information, and representing the results through diagrams and bar charts. These visualizations help identify patterns, compare variables, and support effective decision-making.

The average is obtained by adding all the values in a range and dividing them by the number of observations. By means of the formula =AVERAGE (range). ExcelDemy (2024).

For instance, in this case of film database to generate descriptive statistics such as, average rating, watches, fans, runtime.

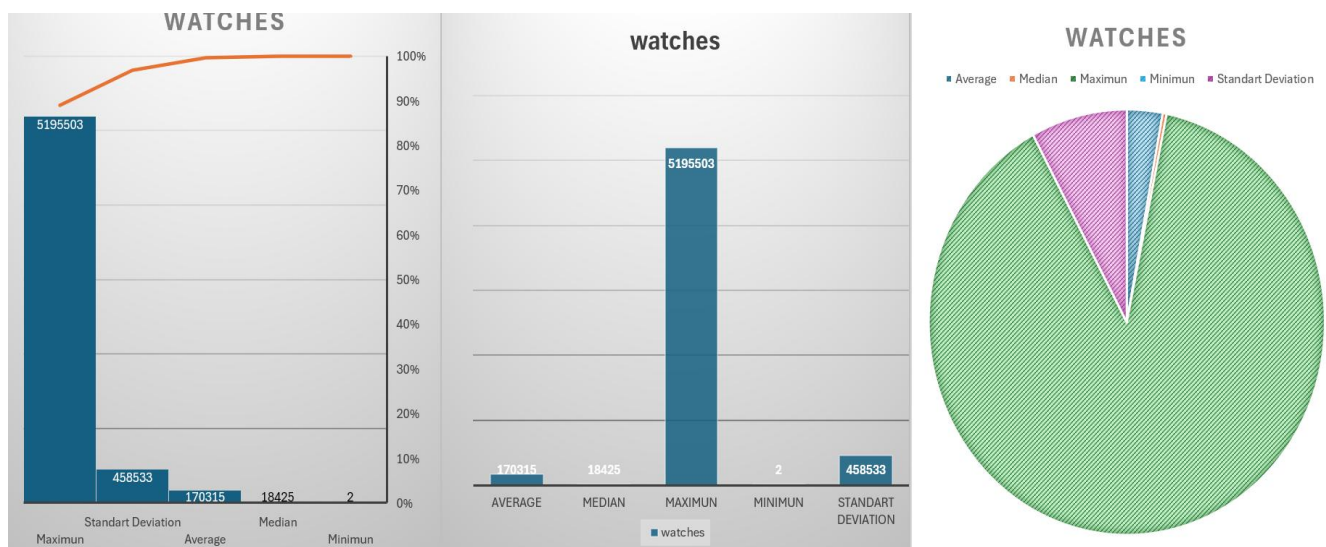
	Average	Median	Maximum	Minimum	Standard Deviation
Fans	1653	24	326000	0	9368
Watches	170315	18425	5195503	2	458533
Runtime	103	99	907	2	38
Average Rating	3	3	4.71	1.08	1

3.1 Descriptive Statistics of Fans



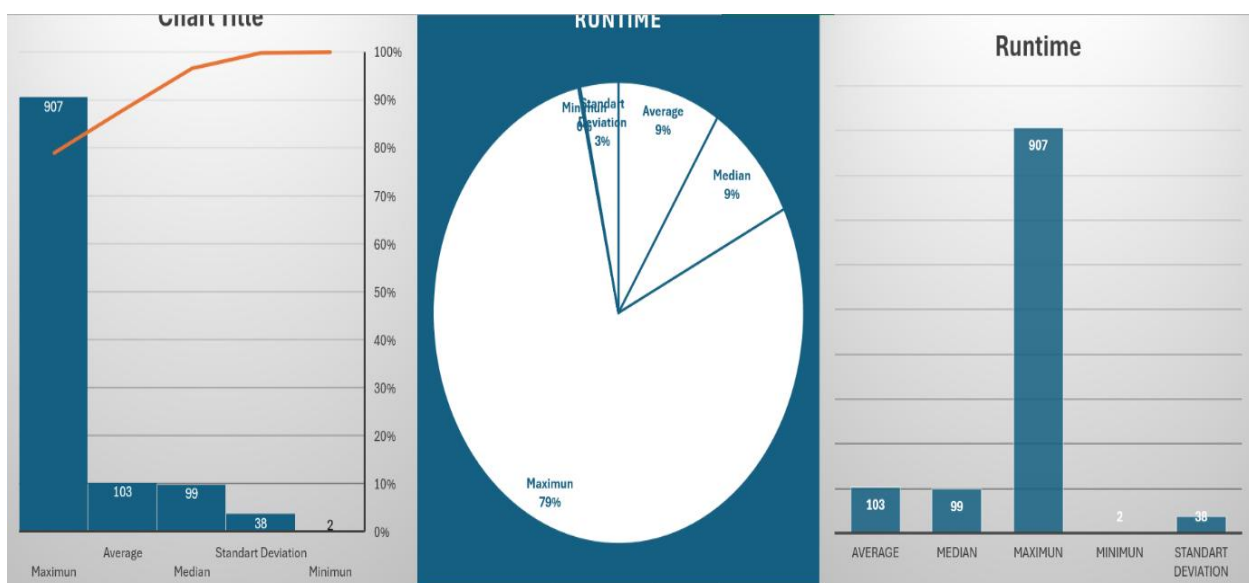
The analysis of the data and graphs reveals that the average number of fans was relatively low (1653). This figure, however, is strongly influenced by the presence of an extreme outlier (326,000), which distorts the overall mean. In contrast, the median is considerably lower (24), reflecting the predominance of films with few or no fans. Furthermore, the very high standard deviation (9388) confirms the existence of substantial variability and dispersion within the distribution of fan counts.

3.2 Descriptive Statistics of Watches



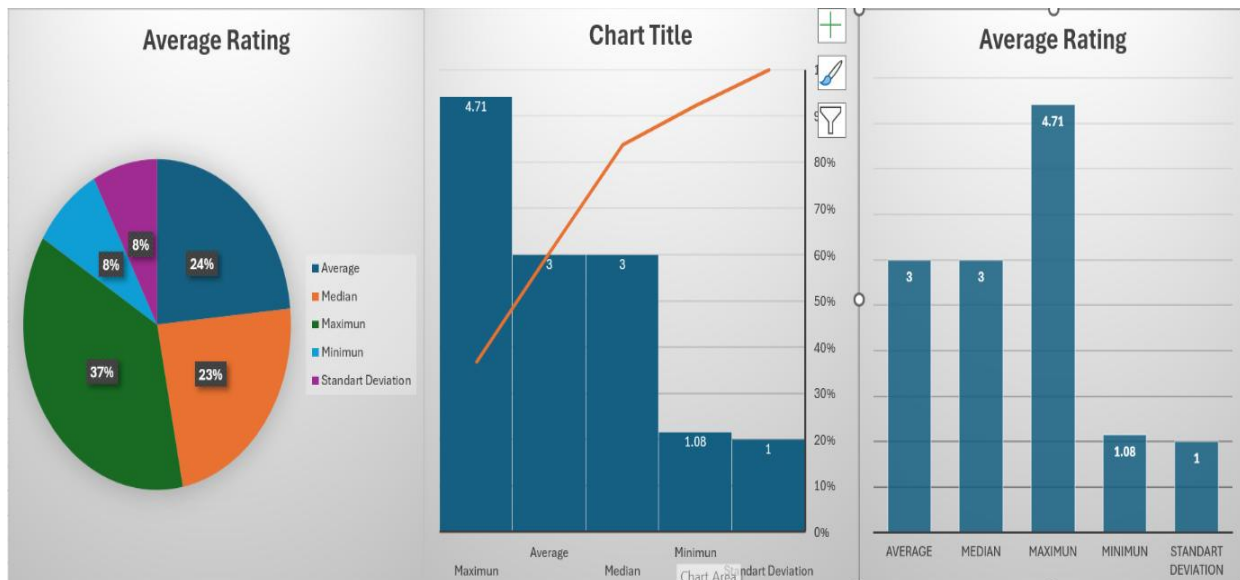
The predictive analysis applied to the variable “watches” indicates that the average number of views is 170,315. This figure represents a relatively modest mean when compared with the extreme maximum of 5,195,503 views and the very low minimum of only 2 views. The median, at 18,425, provides a more representative measure of the typical viewing level. Consequently, the substantial disparity between the maximum and minimum values produces a very large standard deviation of 458,533, confirming the wide dispersion in the distribution of watches.

3.3 Descriptive Statistics of Runtime



The interpretation of this analysis indicates that the average runtime of the films is 103. Although the dataset includes a maximum runtime of 907, and a minimum of only 2, the presence of very short films provokes a downward trend in the overall mean. In contrast, the median runtime of 99, provides a more typical representation of the data. Finally, the standard deviation of 38, confirms a moderate level of variability across the distribution of runtimes.

3.4 Descriptive Statistics of Average Rating



This case of descriptive analysis focuses on the variable “Average Rating.” The results indicate that the mean rating is 3, with values ranging from a maximum of 4.71 to a minimum of 1.08. The median also coincides with the mean at 3, which highlights the central tendency of the dataset. Furthermore, the standard deviation of 1 demonstrates a relatively low level of variability, suggesting that most ratings are clustered closely around the average, despite the observed range.

In conclusion, the results suggest that audience and reception are highly variable, whereas duration and perceived quality tend to be more consistent.

4. Trend Analysis and Visualization

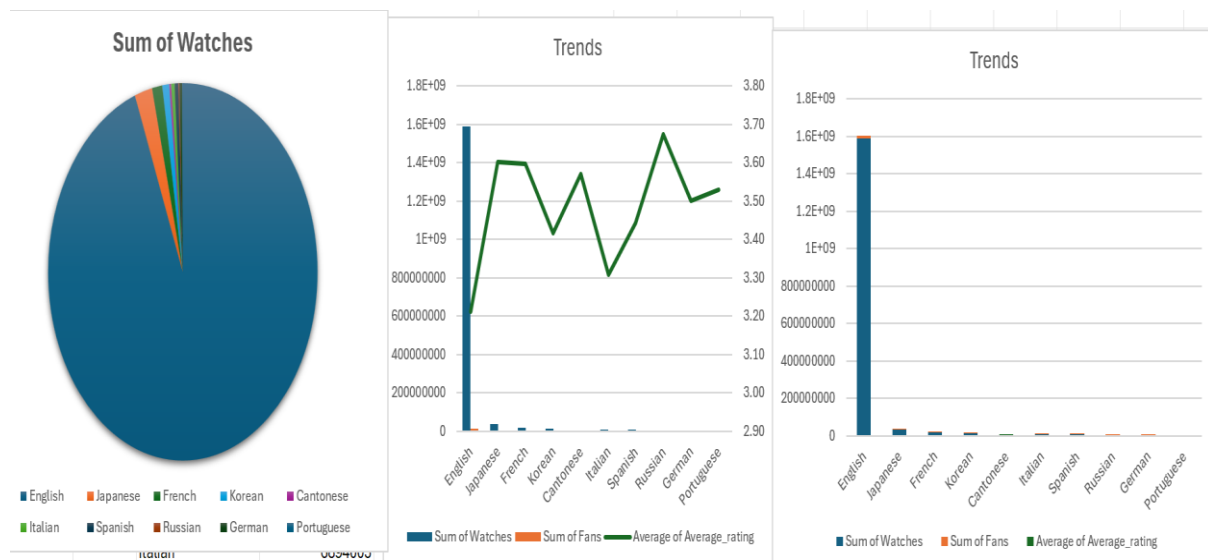
A trend in data analysis is the consistent pattern of variation observed in a variable over time or in relation to another. In descriptive statistics, it is understood as the central tendency of data to display a common behaviour, identified through measures such as mean, median, or mode. In predictive analysis, a trend represents the estimation of future behaviour based on historical

records, modelled with techniques like regression or exponential smoothing. In data visualization, it refers to the line or curve summarizing values' direction. (Cheusheva, 2023)

In this case, to analyse categorical variables such as ratings, views, and fan counts by language in a film database, providing charts to obtain a visual representation that supports effective decision-making.

4.1 Trends correlated with language

Row Labels	Sum of Watches	Sum of Fans	Average of Average_rating
English	1586782495	14307851	3.21
Japanese	35675216	825123	3.60
French	20604729	427112	3.60
Korean	13954038	234449	3.42
Cantonese	3940189	132903	3.57
Italian	6894005	111558	3.31
Spanish	7683190	94738	3.44
Russian	2127522	60736	3.67
German	5313351	52211	3.50
Portuguese	1290508	51792	3.53



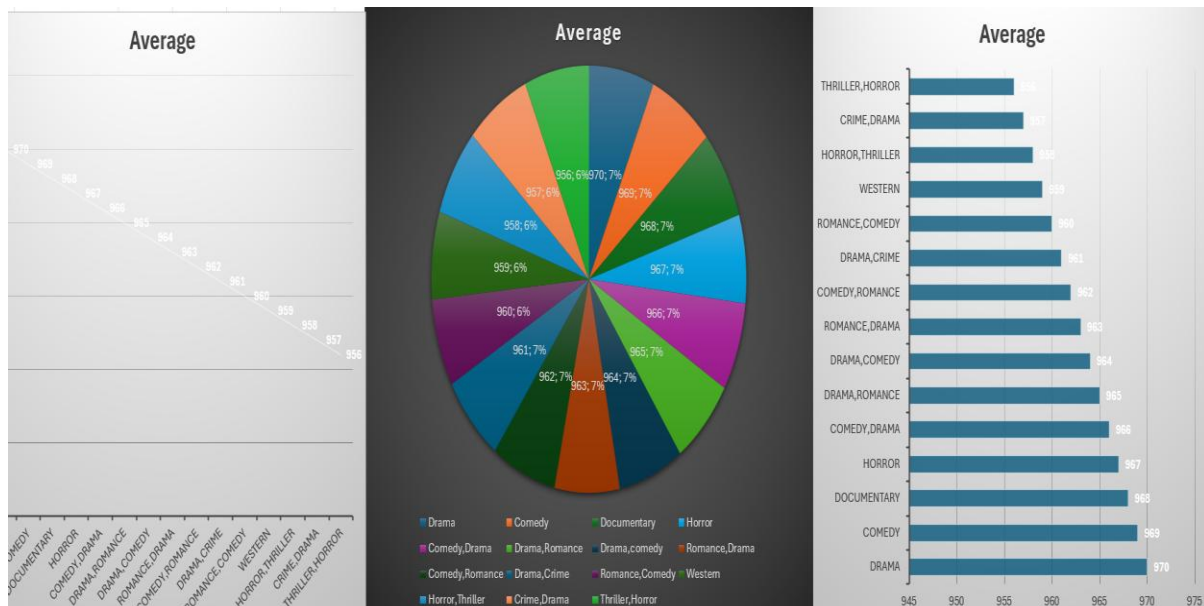
These graphs illustrate the languages with the highest number of views and fans, as well as their average ratings. It can be observed that the English language concentrates the largest amount of watches (1,586,782,495) and fans (14,307,851). However, it obtains a relatively low average rating of 3.21, which suggests that high consumption volume does not necessarily translate into positive evaluation. On the other hand, the languages with the highest average

ratings are Russian (3.67), French (3.60), and Japanese (3.60), although their figures for views and fans remain considerably lower compared to English. In the middle range, with a balance between watches, fans, and ratings, are Spanish (3.44) with 7,683,190 watches and 94,738 fans, followed by Italian (3.31) with 6,894,005 watches and 111,558 fans, and finally German (3.50) with 5,313,351 watches and 52,211 fans. The languages with smaller reach but high audience perception are Portuguese (3.53) with 1,290,508 watches and 51,792 fans, and Cantonese (3.57) with 3,940,189 watches and 132,903 fans. This suggests that, although these are smaller markets, audience satisfaction remains notably high.

In conclusion, the language more popular in watches and fans is English, On the other hand, Russian language is the last of the top is the first place on ranking due to collaboration of the costumers.

4.2 Genres associated with high average ratings

Row Labels	Sum of Average_rating
Drama	970
Comedy	969
Documentary	968
Horror	967
Comedy, Drama	966
Drama, Romance	965
Drama, Comedy	964
Romance, Drama	963
Comedy, Romance	962
Drama, Crime	961
Romance, Comedy	960
Western	959
Horror, Thriller	958
Crime, Drama	957
Thriller, Horror	956



The purpose of this graph is to identify which genres in the film database achieve the highest average ratings and to highlight which genres accumulate greater valuations among the top fifteen. The analysis shows that Drama (970), Comedy (969), and Documentary (968) stand out as the leading individual genres, consistently positioned at the top of the ranking. In contrast, combined genres such as Comedy–Drama (966) and Drama–Romance (965) appear in the middle range, suggesting that these combinations maintain a stable level of quality and audience appreciation. At the lower end, Crime–Drama (957) and Thriller–Horror (956) record weaker performances in terms of average ratings. Additionally, the graph illustrates the proportional distribution of genres. Confirming their relevance in shaping audience preferences. To conclude, Drama is the genre with more followers, as well as in their joint versions with other genres.

5. Statistical Analysis using SPSS

Correlation: The Pearson correlation coefficient measures linear association between two numerical variables. The value of r shows strength and direction: positive means simultaneous increase, negative means inverse. (Obilor, et al, (2018).

T-test: The T-test compares the means of two groups or variables to evaluate whether the observed difference is statistically significant, providing evidence of meaningful variation between them. (Arkkelin, 2014)

Regression: It is a statistical technique that allows the analysis of the relationship between a dependent variable (Y) and one or more independent variables (X), with the purpose of predicting values and evaluating the strength of that relationship. (Lorenzo-Seva, et al, 2010)

5.1 Difference between English and non-English films by Perform t-test.

English language	English average	Non-English	Non-english average				
English	4.18	German	3.57				
English	3.93	Japanese	4.28			English	Non-English
English	3.79	French	3.77		Average	3.21	3.51
English	3.72	French	3.69				
English	4.17	French	4.02				
English	4.25	Spanish	3.77				
English	3.76	Japanese	4.19				
English	3.96	French	3.92				
English	4.15	Italian	4.27				
English	4.38	Cantonese	3.78				
English	3.16	Italian	3.8				
English	4.04	Indonesian	4.02				
English	4.1	Italian	3.52				
English	4.1	Korean	4.38				
English	4.04	Japanese	4.45				
English	4.22	No spoken language	4.02				
English	3.24	French	4.51				
English	2.24	French	4.26				
English	4.42	Russian	4.23				
English	3.08	Russian	4.63				
English	3.71	Japanese	4				
English	3.52	French	3.86				

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Englishaverage	8074	3.2109543889	.52384989467	.00582991857
Nonenglishaverage	1926	3.5099201128	.46117644879	.01050845768

One-Sample Test

Test Value = 0

	t	df	Significance		Mean Difference	95% Confidence Interval of the Difference	
			One-Sided p	Two-Sided p		Lower	Upper
Englishaverage	550.772	8073	<.001	<.001	3.2109543889	3.1995262451	3.2223825327
Nonenglishaverage	334.009	1925	<.001	<.001	3.5099201128	3.4893109562	3.5305292695

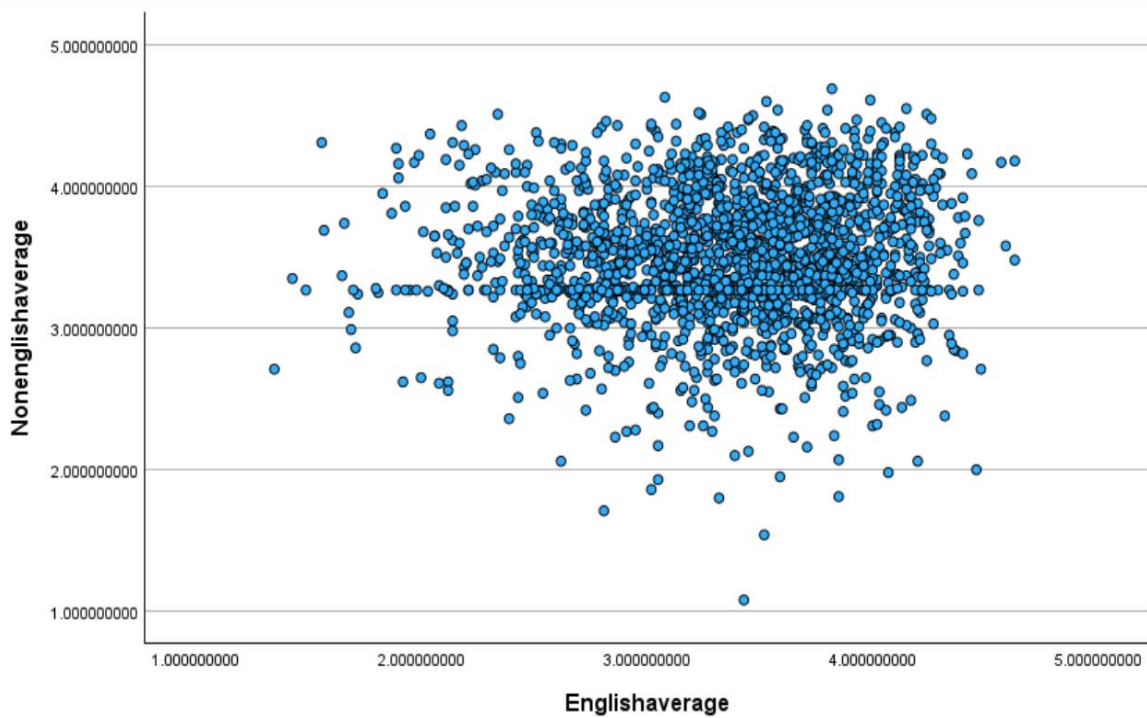
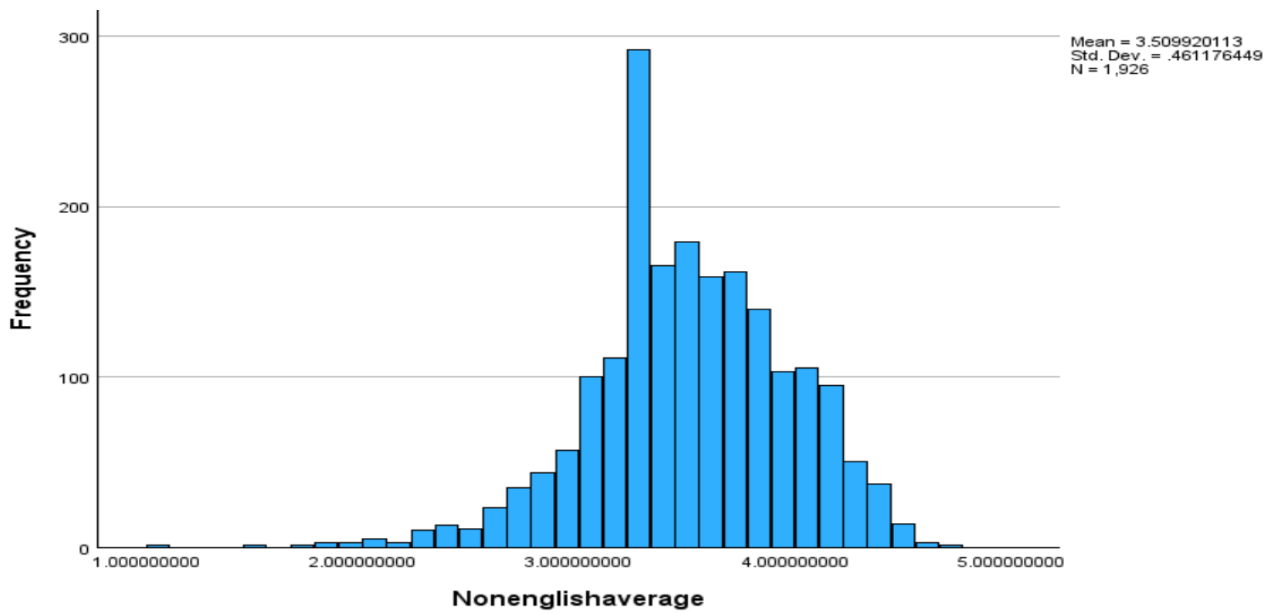
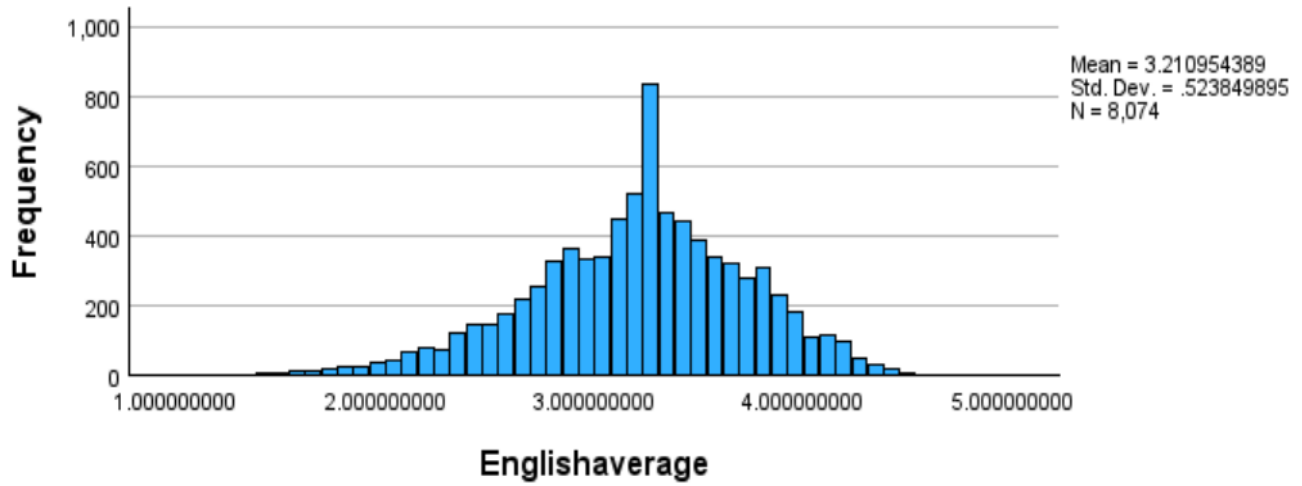
One-Sample Effect Sizes

		Standardizer ^a	Point Estimate	95% Confidence Interval	
				Lower	Upper
Englishaverage	Cohen's d	.52384989467	6.130	6.032	6.227
	Hedges' correction	.52389856779	6.129	6.032	6.226
Nonenglishaverage	Cohen's d	.46117644879	7.611	7.366	7.855
	Hedges' correction	.46135622521	7.608	7.363	7.852

a. The denominator used in estimating the effect sizes.

Cohen's d uses the sample standard deviation.

Hedges' correction uses the sample standard deviation, plus a correction factor.

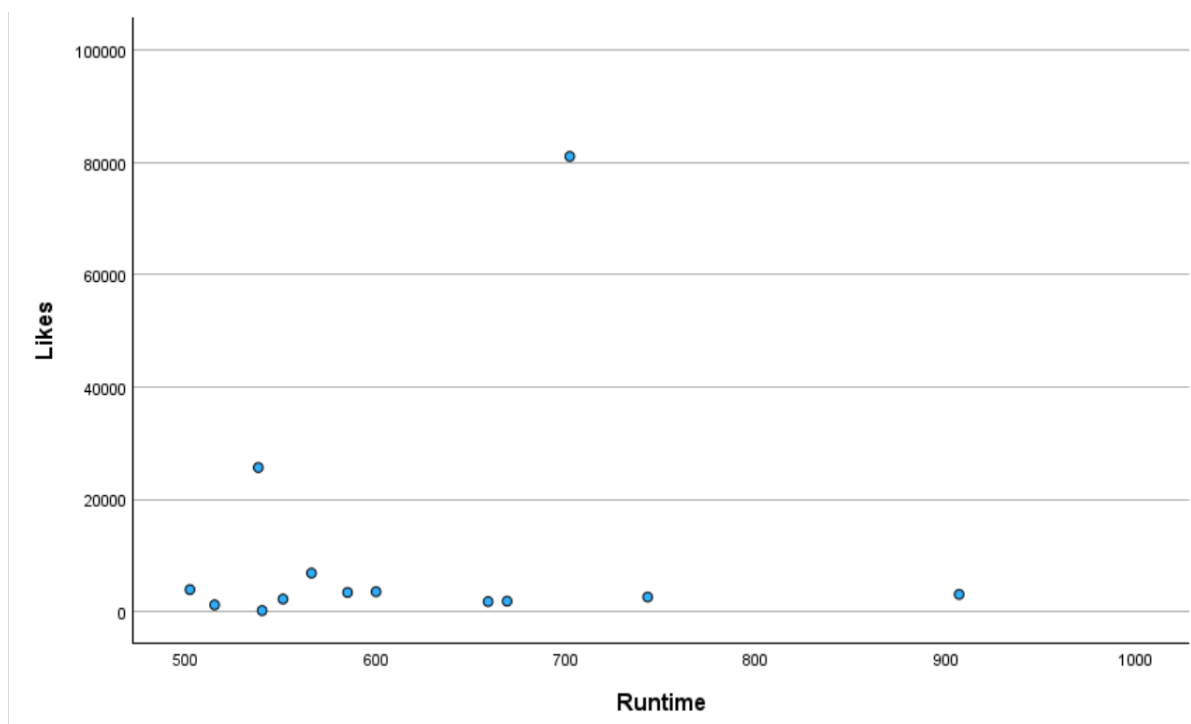


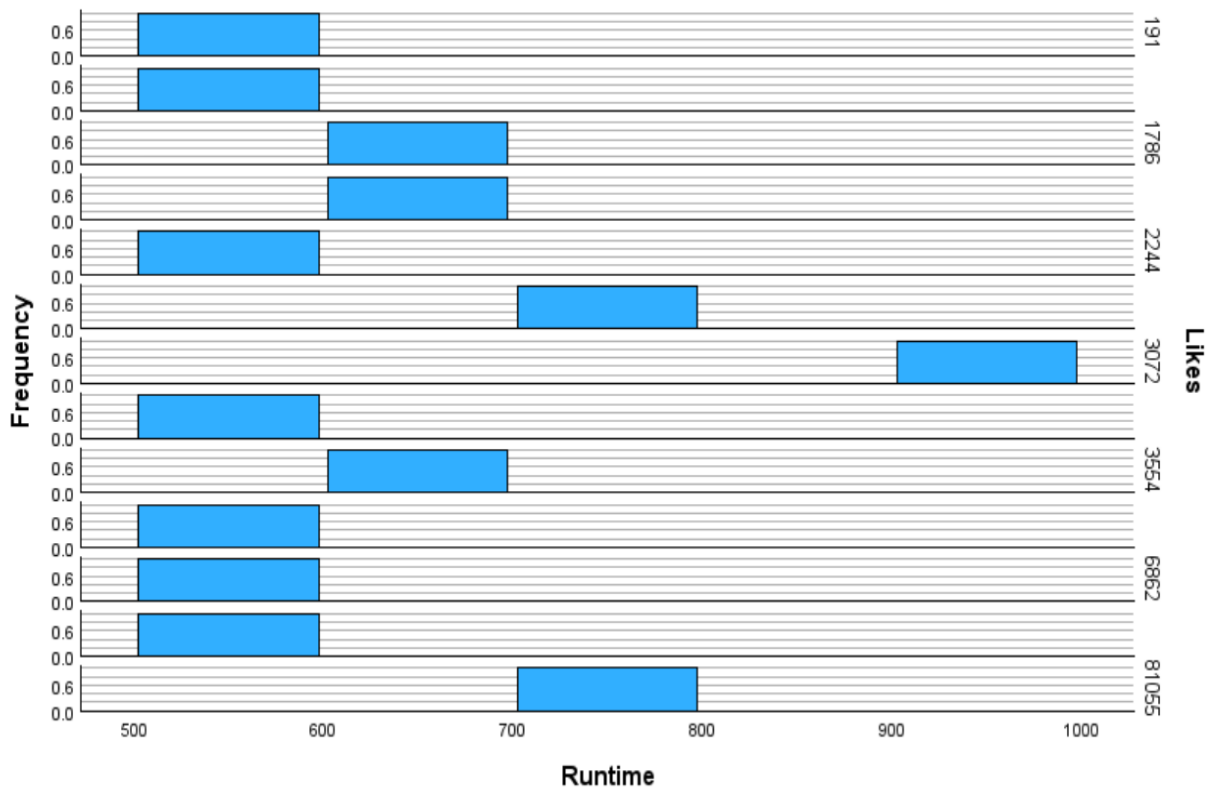
The main objective of this section is to analyse the differences in ranking between films originally produced in English and those in non-English languages. The data reveals a slight advantage for non-English films, which exhibit an average ranking of 3.51, compared to 3.21 for English-language films. The maximum ranking for English films is 3.22, with a minimum of 3.19, while non-English films range from a maximum of 3.53 to a minimum of 3.48. The graph illustrates a minimal difference between the two groups. English-language films display a relatively even distribution across all ratings, whereas non-English films show a pronounced peak at 3.5 followed by a sharp decline. Additionally, the scatter plot suggests that both language categories follow a similar trend.

5.2 Correlations between variables with longer runtimes determine tend to have more watches or likes.

Correlations

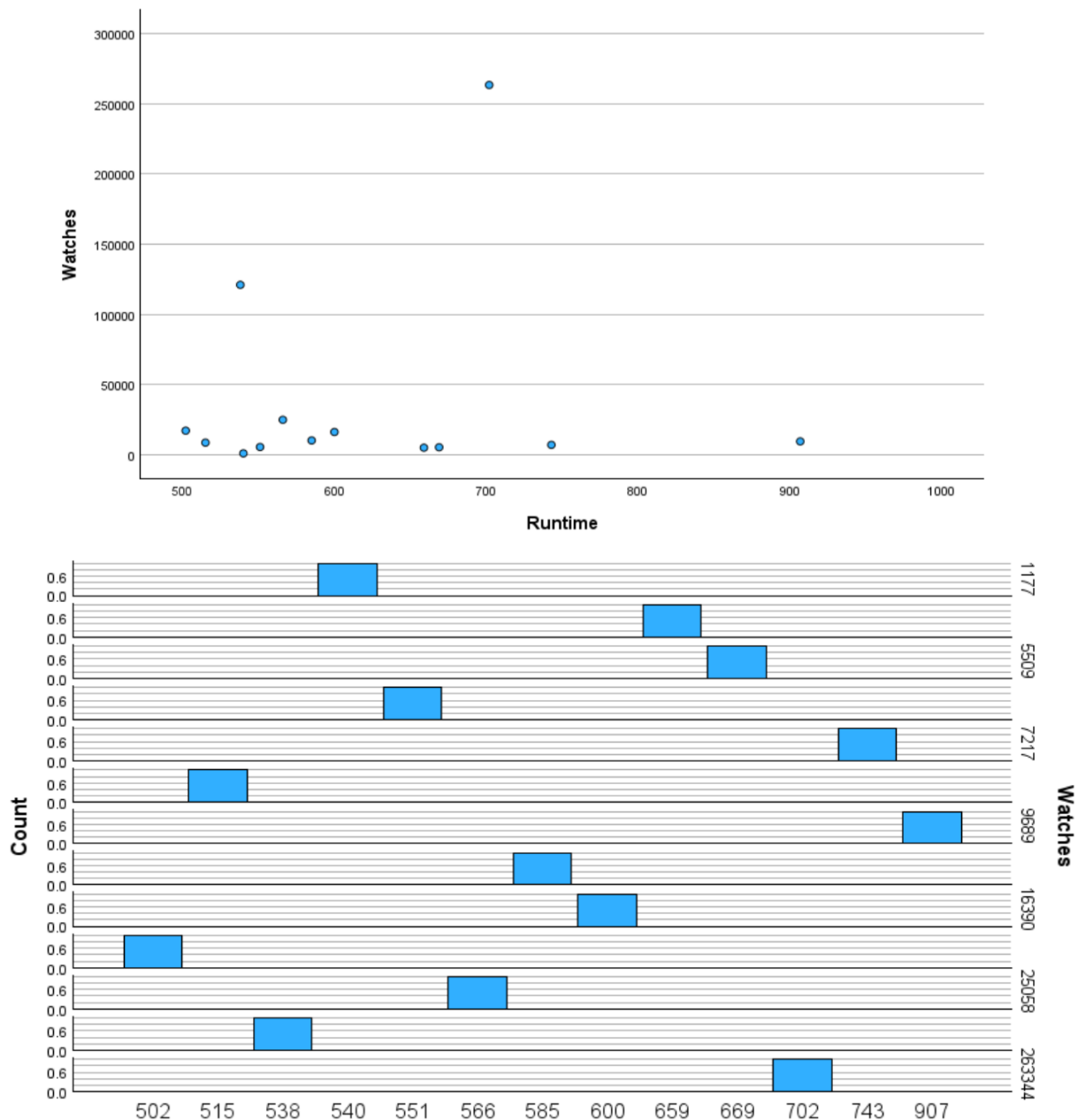
		Runtime	Likes
Runtime	Pearson Correlation	1	.145
	Sig. (2-tailed)		.635
	N	13	13
Likes	Pearson Correlation	.145	1
	Sig. (2-tailed)	.635	
	N	13	13





The correlation between Runtime and Likes is 0.145, indicating a weak positive relationship. The scatter plot shows shorter runtimes sometimes associated with high likes, while longer runtimes, such as around 700 minutes, also appear with large values (e.g., 80,000 likes). However, the significance level (Sig. 2-tailed = 0.635) is far above the conventional 0.05 threshold, meaning the correlation is not statistically significant. The histogram illustrates the distribution, and overall, there is no evidence that Runtime influences the number of Likes.

Correlations			
		Runtime	Watches
Runtime	Pearson Correlation	1	.095
	Sig. (2-tailed)		.758
	N	13	13
Watches	Pearson Correlation	.095	1
	Sig. (2-tailed)	.758	
	N	13	13



The correlation between *Runtime* and *Watches* is 0.095, reflecting a weak relationship between the variables. This indicates no meaningful association between runtime and watches. The scatter plot shows most points clustered at shorter runtimes, with one notable case of approximately 700 minutes and 25,500 watches. The significance value (Sig. 2-tailed = 0.758) is well above the 0.05 threshold, confirming the correlation is not statistically significant. Therefore, there is no evidence that runtime influences watches; variability must be explained by other factors.

5.3 Regression Analysis

Regression

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	Runtime ^b	.	Enter

a. Dependent Variable: Average_rating

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.187 ^a	.035	.035	.51654159946

a. Predictors: (Constant), Runtime

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	96.171	1	96.171	360.441	<.001 ^b
	Residual	2667.619	9998	.267		
	Total	2763.790	9999			

a. Dependent Variable: Average_rating

b. Predictors: (Constant), Runtime

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	3.003	.015		201.241	<.001
	Runtime	.003	.000	.187	18.985	<.001

a. Dependent Variable: Average_rating

These two tables represent the results of a linear regression between runtime and average rating. The model summary shows the correlation coefficient $R = 0.187$, indicating a weak positive relationship, with $R^2 = 0.035$ meaning that 3.5% of the variation in average ratings is explained by movie runtime. The constant coefficient sets the baseline rating at 3.003, while runtime contributes an average increase of 0.003 points for each additional minute of the movie. The standardized beta further confirms the relationship between the two variables. In conclusion, movies with longer runtimes tend to have slightly higher ratings than shorter ones.

6. Recommendations

Strategies and recommendations based on the film database derived from analysis, regression, and average ratings to improve decision-making in Business Intelligence.

Optimization of runtime and genre segmentation: Runtime is highly dispersed; therefore, maintain clear divisions such as short films, feature-length movies, and extended versions. Adjust genres accordingly—for example, horror and comedy benefit from faster pacing, while drama and historical films require greater depth.

Language: Enhance content with multilingual subtitles to increase international ratings, while prioritizing action, sci-fi, and drama genres that demonstrate strong global appeal.

Variables: Expand the dataset to include budget, revenue, release year, and distribution platform. Identifying the main distribution channels will enable predictive analysis for new films

7. Conclusion

Conclusion of the Film Database. The language average is relevant in determining which language ranks at the top level. The English version achieves nearly the same average as the combined values of other languages.

Moreover, runtime has a significant impact on the overall average, since shorter runtimes directly influence the variable as a whole.

Therefore, the grouping of different factors such as runtime, number of fans, and view counts has a substantial effect on the final average, directly shaping the decision-making process.

8. References

Alabi, O. & Bukola, T. (2023) *Introduction to Descriptive Statistics*. In: Kumar, B.S. (ed.) Recent Advances in Biostatistics. IntechOpen. doi:10.5772/intechopen.1002475. Available at: <https://www.intechopen.com/chapters/1141192> [access on 3 December 2025]

Arkkelin, D. (2014). Using SPSS to Understand Research and Data Analysis. *Psychology Curricular Materials. Book*, [online] 1(1). Available at: https://scholar.valpo.edu/cgi/viewcontent.cgi?article=1000&context=psych_oer [Accessed 16 Dec. 2025].

Cheusheva, S. (2023). *Excel TREND function and other ways to do trend analysis*. [online] ablebits.com. Available at: <https://www.ablebits.com/office-addins-blog/excel-trend-function/> [Accessed 7 Dec. 2025].

ExcelDemy (2024). How to Calculate Average in Excel (Ultimate Guide). Available at: <https://www.exceldey.com/learn-excel/calculate/average/> [Accessed 7 Dec. 2025].

Lorenzo-Seva, U., Ferrando, P.J. and Chico, E. (2010). Two SPSS programs for interpreting multiple regression results. *Behavior Research Methods*, 42(1), pp.29–35. doi:<https://doi.org/10.3758/brm.42.1.29>.

Obilor, E.I. and Amadi, E.C., (2018). Test for significance of Pearson's correlation coefficient. *International Journal of Innovative Mathematics, Statistics & Energy Policies*, 6(1), pp.11-23.

Siddiqui, L. (2024). *How to Clean Data in Excel: A Beginner's Guide*. [online] Datacamp.com. Available at: <https://www.datacamp.com/tutorial/data-cleaning-in-excel-a-beginners-guide> [Accessed 1 Dec. 2025].

Sultana, S. (2025). *Data Cleaning in Excel: A Comprehensive Guide for Analysts - ExcelDemy*. [online] ExcelDemy. Available at: <https://www.exceldey.com/data-cleaning-in-excel-a-comprehensive-guide-for-analysts/> [Accessed 1 Dec. 2025].

Tableau (2024). *Data cleaning: the Benefits and Steps to Creating and Using Clean Data*. [online] Tableau Software. Available at: <https://www.tableau.com/learn/articles/what-is-data-cleaning>.