



LCBB5014

BA (Hons) Business Management

Data Handling and Business

Intelligence

LCBB5014

Student Name: Mhaela Cosmina Barabas

Student ID: 2229446

Table of Contents

Introdcution.....	3
Data Preparation, Cleaning and Understanding.....	4
Dataset description.....	4
Data cleaning.....	4
Descriptive Statistics and Summary.....	6
Trend analysis.....	10
Statistical Analysis using SPSS.....	12
T-test.....	12
Correlation.....	14
Regression.....	15
Insights and Recommendations.....	18
References.....	19

List Of Figures:

Figure 1: Average rating issue	4
Figure 2: Run Time variable issue	5
Figure 3: Change data type	5
Figure 4: Variable View	6
Figure 5: Formula Tab for Descriptive Statistics in Excel	6
Figure 6: Numerical Descriptive Statistics	7
Figure 7: Desriptive statistics formula tab in SPSS	7
Figure 8: Descriptive Statistics in SPSS	7
Figure 9: Chart Builder in SPSS	8

Figure 10: Gener Frequency	8
Figure 11: Distribution of Language	9
Figure 12: Box Plot	10
Figure 13: Average Rating by Geners	10
Figure 14: Likes and Fans	11
Figure 15: Encoding Language column	12
Figure 16: Formula tab for t-test	12
Figure 17: T-test	13
Figure 18: Correlation formula Tab in Spss	14
Figure 19: Correlation	14
Figure 20: Regression Formula tab	15
Figure 21: Regression statistics	15
Figure 22: Regression Model Summary	16
Figure 23: Coefficient for Regression Analysis	16
Figure 24: Scatter Plot for Regression model	17

Introdcution

The purpose of this report is to examine viewer behaviour, engagement trends, and trends in the rating of movies, using Letterboxd Movie Classification Dataset which consists of 10,000 movies and 17 variables including movie attributes, ratings, genres, languages, and measures of user interaction, such as watches, likes and fans. The report is aimed at learning how various film properties affect the attention of viewers and their overall performance, and it also reflects competencies in practical data analysis with the use of Excel and SPSS.

Data cleaning starts the analysis process to correct the formatting problem, standardise the numerical fields, and provide accuracy among variables. Descriptive statistics is then produced to summarise the important numerical and categorical data in order to find the patterns behind. Visualisation-based trend analysis investigates the changes in rating, runtime, engagement, and differences between the genres or languages. Lastly, statistical tests, such as correlation analysis, t-tests and regression, are performed to explore the relationship among variables and conclude on meaningful predictors of ratings and engagement. These approaches can be combined to achieve a holistic idea on film performance and audience behaviour.

Data Preparation, Cleaning and Understanding.

Dataset description

The Letterboxd Movie Classification Dataset consists of 10,000 movies and 17 variables of movies, interaction and rating behaviour of viewers. The main characteristics are the title of the film, its director, genres, run time, original language, and the production studios. The watch, appearance of the list, the likes and the number of fans are engagement metrics that measure the use of the platform by the user. The dataset also gives a detailed distribution of the rating in terms of lowest, medium and highest ratings of the stars in terms of counts as well as the total rating. With the help of this structure, it is possible to analyze the patterns of viewing, indicators of popularity, the tendencies of genres, and any elements that impact the audience reception and performance of a movie in various languages and categories.

Data cleaning

The dataset has certain problems, which will be addressed with the help of the following steps:

1	SN	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
		Film_title	Director	Average	Genres	Runtime	Original	Descrip	Studios	Watche	List_appe	Likes	Fans	Lowest	Medium	Highest	Total_r	gs	
507	505	Came a Hc lan M	3.2685351	[Drama', 'C	101	English	Set in post-['New Zeal	260	398	36	1	3	27	3	112				
513	511	Shaker Rui Bruce Mor	3.2685351	[Action', 'C	90	English	With the ac['Mirage-A	254	658	47	2	4	47	4	152				
518	516	The Silent (Yvonne Ma	3.2685351	[Drama', 'F	92	English	A baby is w['The Gibsc	110	420	18	1	1	8	4	39				
748	746	American Michael C	3.2685351	[Comedy',	97	English	Somewhere ['Harmony	284	502	39	0	22	25	7	174				
832	830	The Lawless Jon Hess	3.2685351	[Action', 'S	81	English	Action-adv Unknown	120	485	11	0	8	9	0	70				
893	891	Tex and the Duccio Te	3.2685351	[Adventure	104	Italian	Based on t['RAI', 'SAC	275	495	16	0	16	37	3	193				
1252	1250	Yellow Nick Cass	3.2685351	[Drama']	109	English	A young su Unknown	196	983	32	2	3	31	5	111				
1456	1454	Deathmas Richard Fr	3.2685351	[Mystery',	103	English	Haunted by ['Art Theatr	139	506	18	0	5	18	0	79				
1513	1511	Shopping Louis Suth	3.2685351	[Drama']	98	English	Living in Ne['Carver Fil	156	374	25	1	1	25	2	89				
1605	1603	The Highw Douglas H	3.2685351	[TV Movie']	91	English	A governm Unknown	162	332	11	0	3	22	3	67				
1606	1604	Under the James Sba	3.2685351	[Action', 'C	86	English	A St. Louis ['Marquis F	89	408	13	0	3	14	1	57				
1658	1656	Ten Zan - U Ferdinand	3.2685351	[Drama', 'J	84	English	The story ir ['Amerinda	138	380	9	0	18	12	1	98				
1659	1657	Thunder III Fabrizio D	3.2685351	[Drama', 'J	90	Italian	They took f ['Fulvia Filr	190	536	31	0	8	32	4	137				
1700	1698	Outlaw Rid Tony Hust	3.2685351	[Action']	86	English	An outlaw i Unknown	41	354	2	0	4	5	1	21				
1701	1699	The Snake Piero Vivar	3.2685351	[Horror', 'J	94	Italian	A beautiful ['Filmes Ve	218	866	30	0	5	34	2	146				
2089	2087	Beyond the Matthias H	3.2685351	[Thriller', 'I	93	English	An English ['Hammer	263	658	25	2	40	15	5	178				
2272	2270	Baywatch Douglas Sc	3.2685351	[TV Movie']	90	English	The world's ['Tower 12	368	372	36	0	35	34	13	194				
2277	2275	Leapin' Ler Ted Nicola	3.2685351	[Fantasy']	84	English	A man tries ['Moonbea	195	456	29	0	9	28	4	114				
2286	2284	The Incred George Er	3.2685351	[Adventure	97	English	In this glob Unknown	74	277	6	0	2	8	2	42				
2324	2322	Blue Deser Bradley Ba	3.2685351	[Thriller', 'I	97	English	After survi Unknown	265	496	28	0	8	35	6	169				
2372	2370	Domenica Wilma Lab	3.2685351	[Drama']	95	Italian	Police insg ['Sidecar F	23	270	2	0	2	5	1	10				
2384	2382	Terror on T Robert Wo	3.2685351	[Documen	85	English	A video sto ['continent	208	246	33	0	4	26	3	109				
2387	2385	Sunshine Jacky Lee	3.2685351	[Comedy']	16	English	A geeky yoi ['TheBoyW	93	356	16	0	3	16	0	52				
2393	2391	Cat in the (Tony Zarim	3.2685351	[Thriller']	98	English	A young mi ['Shapiro E	163	564	26	0	8	17	4	104				
2465	2463	Brotherly L. J. Lee Thor	3.2685351	[Drama']	107	English	Having left ['Keep Film	165	401	29	0	1	27	3	97				

Figure 1: Average rating issue

Issue is resolved by changing this value as number form.

1	SN	Film_title	Director	Average	Genres	Runtime	Original	Description	Studios	Watches	List_appearances	Likes	Fans	Lowest	Medium	Highest	Total Ratings
4089	4087	S de Sofia	Gabe Gabr	3.27	['Crime']	103.16833466773419		Unknown		26	223	3	0	0	2	0	6
6359	6357	Dance Car Bertie, Ber		3.27	['Comedy']	103.16833 English		When Hun [YouTube I		302	357	20	1	18	17	5	150
6515	6513	State of Plc David Yate		4.01	['Mystery']	103.16833 English		The murde [Unknown		3773	1017	724	1	5	197	317	1979
9107	9106	Amend: Th Kenny Leor		3.67	['Documentary']	103.16833 English		Will Smith [The Docu		1448	617	315	1	5	80	167	790
9168	9167	Augustus: Roger Your		3.27	['Drama']	103.16833 English		Old and ill, [Lux Vide]		300	411	36	1	15	42	8	189
9512	9511	Highlander Chad Stah		3.27	['Action']	103.16833 English		An immort [Summit E		254	813	73	0	0	0	0	0
9652	9651	Getting Pei Ron Burrus		3.27	['Drama']	103.16833 English		Young prof [The AsyLu		12	145	2	0	0	0	2	9
9911	9911	The Ballad Edward Be		3.27	['Drama']	103.16833 English		A high-stal [Good Che		30	2136	16	0	0	0	0	0
9976	9976	The Memoi Reed Mora		3.27	['Thriller']	103.16833 English		On an unnn [Picture Fi		122	2169	66	0	0	0	0	0
10002																	
10003																	
10004																	
10005																	
10006																	

Figure 2: Run Time variable issue

This issue is resolved by replacing the value from the correct number format.

File Home Insert Page Layout Formulas Data Review View Automate Developer Help Power Pivot

Clipboard Font Alignment

123 General No specific format

12 Number Total_ratings

Currency Total_ratings

Accounting Total_ratings

Short Date Total_ratings

Long Date Total_ratings

Time Total_ratings

Percentage Total_ratings

Fraction Total_ratings

Scientific Total_ratings

More Number Formats...

SN	Film_title	Director	Average	Genres	Runtime	Original	Description	Studios	Watches	List_appearances	Likes	Fans	Lowest	Medium	Highest	Total Ratings	
1	0	The Fan	Eckhart Sc	3.57	['Horror', 'D	92	German	The hero-w [Barbara M		1284							
2	1	Mad Max: f George Mil		4.18	['Adventure	121	English	An apocal [Warner B		248872							
3	2	Suspiria	Dario Arge	3.93	['Horror']	99	English	An Americ [Seda Spe		61817							
4	3	Lost in Trar Sofia Copp		3.79	['Drama', 'C	102	English	Two lost so [American		159619							
5	4	Akira	Katsuhiro C	4.28	['Animation	124	Japanese	A secret m [MBS', 'Sur		87306							
6	5	Things to C Mia Hanse		3.77	['Drama']	102	French	Nathalie te [ARTE Frar		3477							
7	6	Big Trouble John Carpe		3.72	['Comedy']	100	English	Truck drive [TAFT Ente		28387							
8	7	Star Wars	George Luk	4.17	['Adventure	121	English	Princess Li [Lucasfilm		301432							
9	8	The Third P Carol Reec		4.25	['Mystery', 'I	105	English	In postwar [London Fi		22816							
10	9	The Big Sic Michael Sh		3.76	['Drama', 'C	120	English	Pakistan-b [FilmNatio		38220							
11	10	Jodorowsk Frank Pavi		3.96	['Document	90	English	Shot in Fra [Snowfort		8538							
12	11	Get Out	Jordan Pee	4.15	['Horror', 'H	104	English	Chris and f [Monkeyp		437600							
13	12	Lawrence David Lear		4.38	['Adventure	228	English	The story o [Horizon P		36974							
14	13	Free Fire	Ben Wheat	3.16	['Action', 'C	90	English	Set in Bost [Rook Film		10984							
15	14	Sing Street John Carne		4.04	['Romance	105	English	A boy grow [Likely Sto		39091							
16	15	Up	Pete Docte	4.10	['Animation	96	English	Carl Fredri [Pixar']		392221							
17	16	Blade Runn Ridley Scot		4.10	['Thriller', 'I	118	English	In the smoj [Shaw Bro		165884							
18	17	Annie Hall Woody Alle		4.04	['Drama', 'F	93	English	New York c [United Ar		45423							
19	18	The Truma Peter Weir		4.22	['Comedy', 'I	103	English	Every seco [Paramou		4382819	361732	1777004	71000	5764	231650	930042	2926289
20	19	Wonder W Patty Jenki		3.24	['Action', 'F	141	English	An Amazor [Atlas Ente		1721440	147437	320485	1600	25306	239158	70337	946010
21	20	Plan 9 from Edward D.		2.24	['Science F	79	English	In Californi [Reynolds		61992	20589	11196	121	6724	4949	2272	35588
22	21	Whiplash Damien Ch		4.42	['Drama', 'H	107	English	Under the i [Bold Film		3925603	481379	1848733	154000	7390	146378	1272739	2963756
23	22	Weird Scie John Hugh		3.08	['Comedy', 'I	93	English	Two unpop [Universal		141350	26516	26420	438	1942	21771	4897	76048
24	23	Raw	Julia Ducoi	3.69	['Drama', 'H	99	French	In Justineã [Wild Bunc		519561	150570	150137	6600	5724	59758	41195	375195
25	24	Army of Da Sam Raimi		3.71	['Comedy', 'I	81	English	Ash, a hani [Renaissa		363502	108548	123314	3500	1877	39826	34201	251175

Figure 3: Change data type

Change the datatypes of the variables “Watches, run time, average, likes, ratings, lowest rating, highest ratings, List_appearances, Fans, medium and total ratings”.

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	SN	Numeric	5	0		None	None	12	Right	Scale	Input
2	Film_title	String	84	0		None	None	50	Left	Nominal	Input
3	Director	String	47	0		None	None	47	Left	Nominal	Input
4	Average_rati...	Numeric	19	15		None	None	15	Right	Scale	Input
5	Genres	String	195	0		None	None	50	Left	Nominal	Input
6	Runtime	Numeric	16	12		None	None	13	Right	Scale	Input
7	Original_lan...	String	32	0		None	None	32	Left	Nominal	Input
8	Description	String	1038	0		None	None	50	Left	Nominal	Input
9	Studios	String	568	0		None	None	50	Left	Nominal	Input
10	Watches	Numeric	7	0		None	None	12	Right	Scale	Input
11	List_appear...	Numeric	6	0		None	None	12	Right	Scale	Input
12	Likes	Numeric	7	0		None	None	12	Right	Scale	Input
13	Fans	Numeric	6	0		None	None	12	Right	Scale	Input
14	Lowest	Numeric	6	0		None	None	12	Right	Scale	Input
15	Medium	Numeric	6	0		None	None	12	Right	Scale	Input
16	Highest	Numeric	7	0		None	None	12	Right	Scale	Input
17	Total_ratings	Numeric	7	0		None	None	12	Right	Scale	Input

Figure 4: Variable View

No other issue is in the dataset, no duplicate values.

Descriptive Statistics and Summary

Figure 5: Formula Tab for Descriptive Statistics in Excel

Description Statistics	Average_rating	Runtime	Watches	List_appearances	Likes	Fans	Lowest	Medium	Highest	Total_ratings
Mean	3.268535	103.1694	170314.8	26868.8	43734.79	1652.55	1837.32	17887.08	18171.58	106392.3
Standard Error	0.005257	0.380707	4585.329	567.5133	1515.314	93.68258	56.17492	437.1312	832.346	2947.343
Median	3.268535	99	18425	6503	3394	24	181	2422.5	572	11823
Mode	3.268535	90	226	379	16	0	4	2	0	105
Standard Deviation	0.525744	38.07074	458532.9	56751.33	151531.4	9368.258	5617.492	43713.12	83234.6	294734.3
Sample Variance	0.276407	1449.381	2.1E+11	3.22E+09	2.3E+10	87764260	31556221	1.91E+09	6.93E+09	8.69E+10
Kurtosis	0.394023	72.7118	32.5394	27.64707	74.56651	324.8083	178.013	30.76807	145.5768	41.16012
Skewness	-0.4078	5.649689	5.130309	4.524035	7.497859	14.62154	9.850785	4.80822	10.22969	5.626936
Range	3.63	905	5195501	803637	2730813	326000	176637	629879	2010363	4091037
Minimum	1.08	2	2	33	0	0	0	0	0	0
Maximum	4.71	907	5195503	803670	2730813	326000	176637	629879	2010363	4091037
Sum	32685.35	1031694	1.7E+09	2.69E+08	4.37E+08	16525497	18373196	1.79E+08	1.82E+08	1.06E+09
Count	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000

Figure 6: Numerical Descriptive Statistics

Figure 7: Descriptive statistics formula tab in SPSS

Descriptives

Descriptive Statistics											
	N Statistic	Range Statistic	Minimum Statistic	Maximum Statistic	Mean Statistic	Std. Deviation Statistic	Variance Statistic	Skewness Statistic	Std. Error	Kurtosis Statistic	Std. Error
Average_rating	10000	3.630000000	1.080000000	4.710000000	3.268535187	.5257438723	.276	-.408	.024	.394	.049
Runtime	10000	905.0000000	2.000000000	907.0000000	103.1693515	38.07073527	1449.381	5.650	.024	72.712	.049
Watches	10000	5195501	2	5195503	170314.82	458532.917	2.103E+11	5.130	.024	32.539	.049
List_appearances	10000	803637	33	803670	26868.80	56751.329	3220713309	4.524	.024	27.647	.049
Likes	10000	2730813	0	2730813	43734.79	151531.431	2.296E+10	7.498	.024	74.567	.049
Fans	10000	326000	0	326000	1652.55	9368.258	87764260.16	14.622	.024	324.808	.049
Lowest	10000	176637	0	176637	1837.32	5617.492	31556221.08	9.851	.024	178.013	.049
Medium	10000	629879	0	629879	17887.08	43713.123	1910837130	4.808	.024	30.768	.049
Highest	10000	2010363	0	2010363	18171.58	83234.601	6927998828	10.230	.024	145.577	.049
Total_ratings	10000	4091037	0	4091037	106392.30	294734.320	8.687E+10	5.627	.024	41.160	.049
Valid N (listwise)	10000										

Figure 8: Descriptive Statistics in SPSS

The descriptive statistics display the central tendency and variability of important movie metrics of 10,000 movies (Excel-easy.com, 2025). The mean score on movies is 3.27 (low variance $SD = 0.53$), and the average length of the movie is 103 minutes (large $SD = 38$). The range and skewness of watches, likes and the number of fans are incredibly large, which means that a few popular movies that the users are highly interested in have a dominant position. The variables of rating distribution variables (Lowest, Medium, Highest, Total ratings) also represent a high variance because of the difference in interactions with the audience. Generally, the data has a great variety of film performance, and the degree of engagement is highly skewed (Kent.edu, 2018).

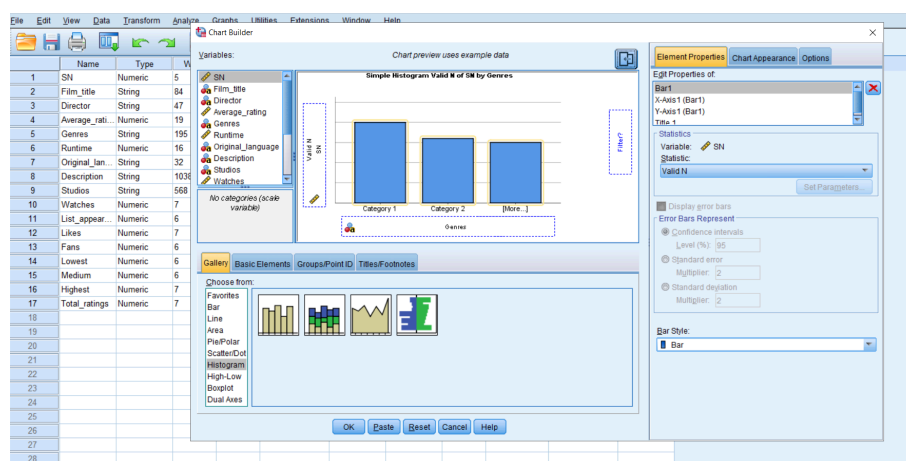


Figure 9: Chart Builder in SPSS

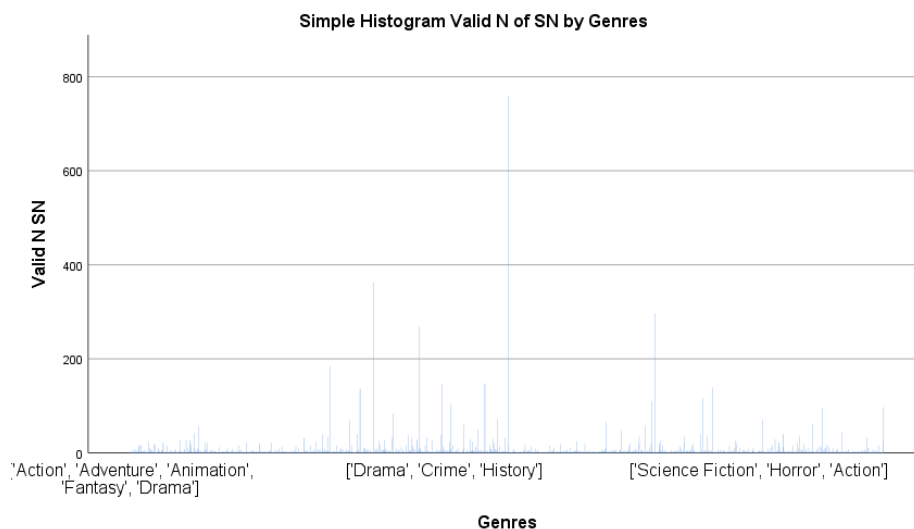


Figure 10: Gener Frequency

The histogram illustrates the distribution of film counts for genres, with the majority having relatively low counts, while a smaller number appear relatively frequently. The few genre combinations dominate, reflecting their over-representation in the dataset as a whole. In general, the distribution of the film counts is highly uneven, and there is notably strong variability in the number of films in each genre.

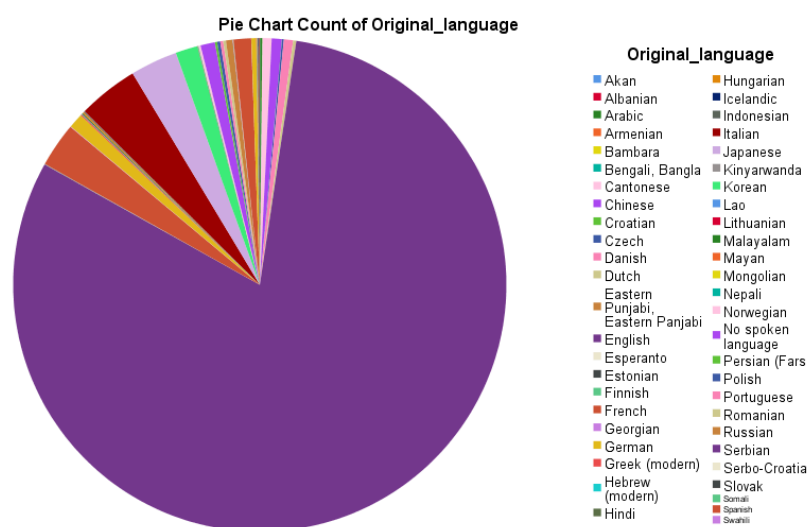


Figure 11: Distribution of Language

According to the pie chart, English dominates as the source language of the majority of films with the huge percentage of the distribution. The other languages are just represented in small proportions. This raises a high level of imbalance where the English-language films constitute the bulk of the data (Baylor.edu, 2025).

The main descriptive statistics were calculated in Excel and SPSS to present the data on numerical and categorical variables of the dataset. The statistics table of descriptive statistics of the SPSS draws attention to the central tendency and dispersion of major metrics of 10,000 films. The means of the movie rating are around 3.27, and the variability is low, but runtime is more dispersed. The variables of engagement like watches, likes, and number of fans have very huge ranges and skewness, meaning that users are concentrated on a few highly popular movies. The imbalance is further presented by excel boxplots which depict a lot of outliers with high values.

The tables and charts of SPSS were used to analyse categorical variables. Histogram of genre distribution indicates that not only do a few genres have high frequency of occurrences but there are numerous genres that are rare.

Trend analysis

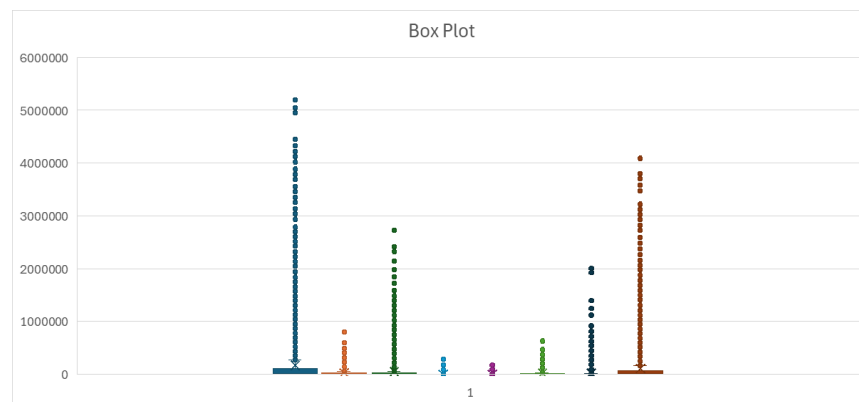


Figure 12: Box Plot

The box plot shows extreme variability across engagement metrics, with numerous high-value outliers for watches, likes and total ratings. Most data points cluster near the lower end, indicating that only a small number of films receive exceptionally high engagement while the majority remain relatively low in user interactions.

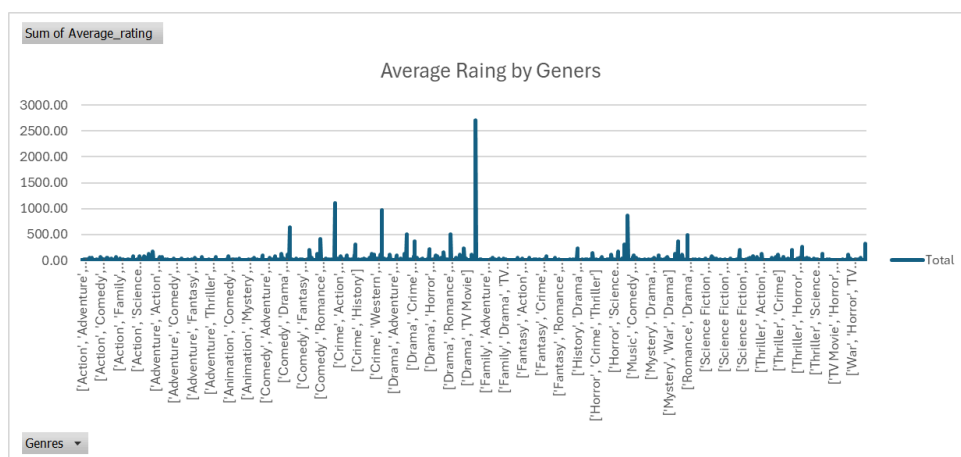


Figure 13: Average Rating by Genres

The chart shows total average ratings across different genres, revealing substantial variation. Some genres, especially drama-related combinations, accumulate much higher total ratings due to their larger number of films. Less common genres display lower totals, indicating fewer entries rather than lower quality. The pattern reflects genre popularity and representation in the dataset (Optimize Smart, 2025).

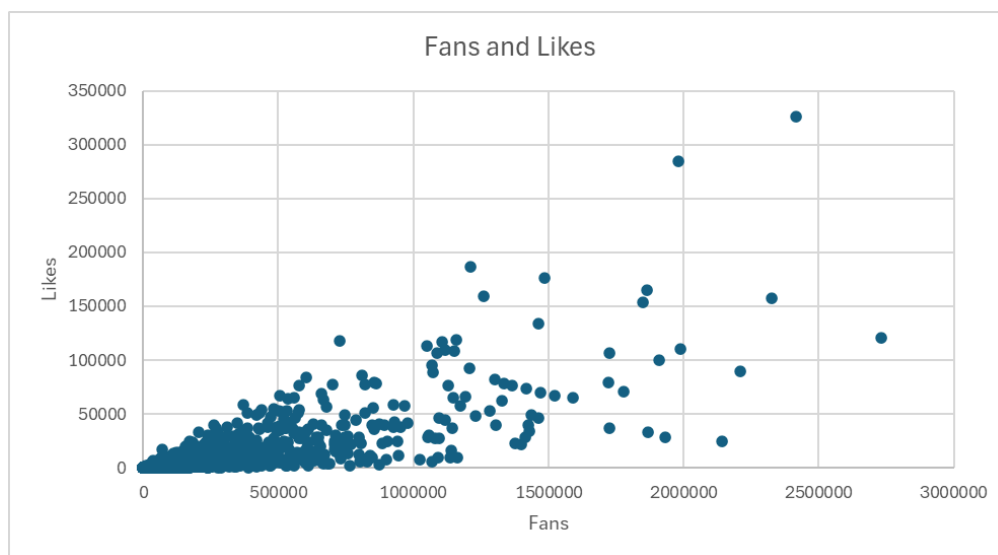


Figure 14: Likes and Fans

The scattered plot indicates that there is a positive correlation between fans and likes. Movies that have many fans tend to have more likes whereas the difference increases towards maximum

values, which means that there is a degree of variance among movies with many followers. The trend indicates that fanbase magnitude is a good indicator of audience involvement.

The major tendencies in ratings, views, and engagement by genre and language were explored and visualised with the help of Excel. The box plot analysis has shown that the measures of viewers interaction like watches, likes, and overall rating are characterized by very high variability, with few movies having an unrealistically high importance. This underlines the fact that the engagement is so skewed, and a few movies can produce a phenomenal response in the audience.

In order to examine the patterns at the genre level, a bar graph of the mean rating per genre revealed significant differences among different categories. Drama, crime or character-driven genre-related genres scored higher on overall ratings primarily due to their prevalence in the data set. Genres that were less popular had lower totals, since their numbers of entries were lower. This implies that those genres that have a wider representation are more likely to get more aggregate ratings, and thus it demonstrates the trends of popularity on the platform.

The scatter plot of fans and likes indicated that there is a high positive correlation with the number of fans indicating that a movie with a bigger number of fans always gets more likes. But the diffusion becomes broader in case of popular movies and it is possible to assume that fans have different approaches to other titles. In general, the trends in engagement, the popularity of genres, and rating behaviour in the entire dataset can be identified through these Excel visualisations.

Statistical Analysis using SPSS

T-test

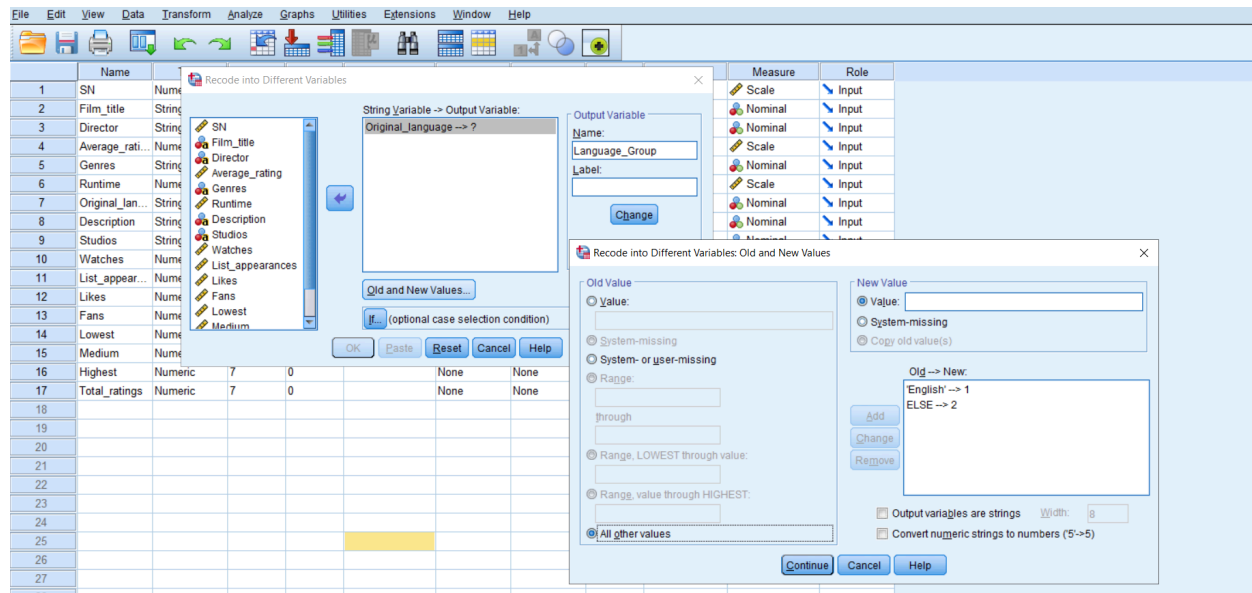


Figure 15: Encoding Language column

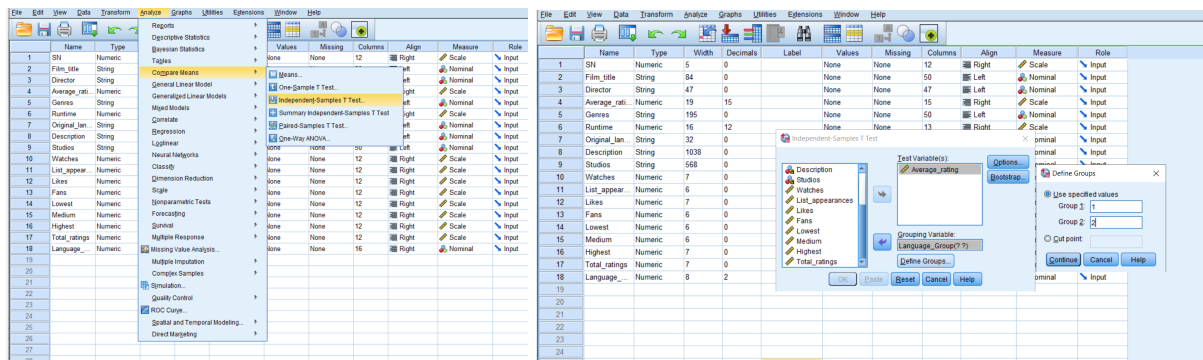


Figure 16: Formula tab for t-test

Hypothesis Statement

H₀: There is no significant difference in average ratings between English and non-English films.

H₁: There is a significant difference in average ratings between English and non-English films.

T-Test

Group Statistics					
	Language_Group	N	Mean	Std. Deviation	Std. Error Mean
Average_rating	1.00	8074	3.210954389	.5238498947	.0058299186
	2.00	1926	3.509920113	.4611764488	.0105084577

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 Lower Upper
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 17: T-test

The independent samples t-test indicates a statistically significant difference in average ratings between the two groups ($t = -23.009$, $p < .001$). Non-English films have a higher average rating ($M = 3.51$) compared to English films ($M = 3.21$). Levene's test is significant ($p < .001$) so there is evidence of unequal variances, but both rows confirm the same outcome. Therefore, we reject the null hypothesis and conclude that language has meaningful effects on audience ratings.

Correlation

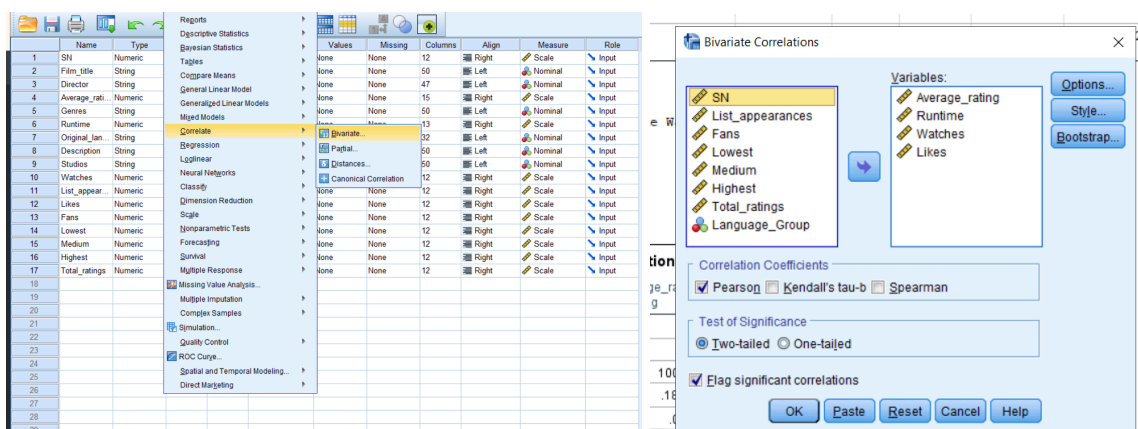


Figure 18: Correlation formula Tab in Spss

Correlations

Correlations					
		Average_rating	Runtime	Watches	Likes
Average_rating	Pearson Correlation	1	.187**	.249**	.290**
	Sig. (2-tailed)		.000	.000	.000
	N	10000	10000	10000	10000
Runtime	Pearson Correlation	.187**	1	.146**	.129**
	Sig. (2-tailed)	.000		.000	.000
	N	10000	10000	10000	10000
Watches	Pearson Correlation	.249**	.146**	1	.951**
	Sig. (2-tailed)	.000	.000		.000
	N	10000	10000	10000	10000
Likes	Pearson Correlation	.290**	.129**	.951**	1
	Sig. (2-tailed)	.000	.000	.000	
	N	10000	10000	10000	10000
**. Correlation is significant at the 0.01 level (2-tailed).					

Figure 19: Correlation

The results of correlation indicate that the positive relationship between runtime, watches, and likes is statistically significant (Laerd.com, 2018). The positive correlation between runtime and watches ($r = .146$, $p < .001$) is weak although positive, which means that longer run time movies are slightly more likely to have watches however, the relationship is not strong. Runtime has also a weak positive correlation with likes ($r = .129$, $p < .001$) implying that longer movies tend to get more likes, but once again there is low strength. Watch and likes have the highest correlation in the table ($r = .951$) and it implies that widely watched movies also get a lot of likes. In general, the effects of the runtime on engagement are slight, whereas the relationships between watches and likes are very strong.

Regression

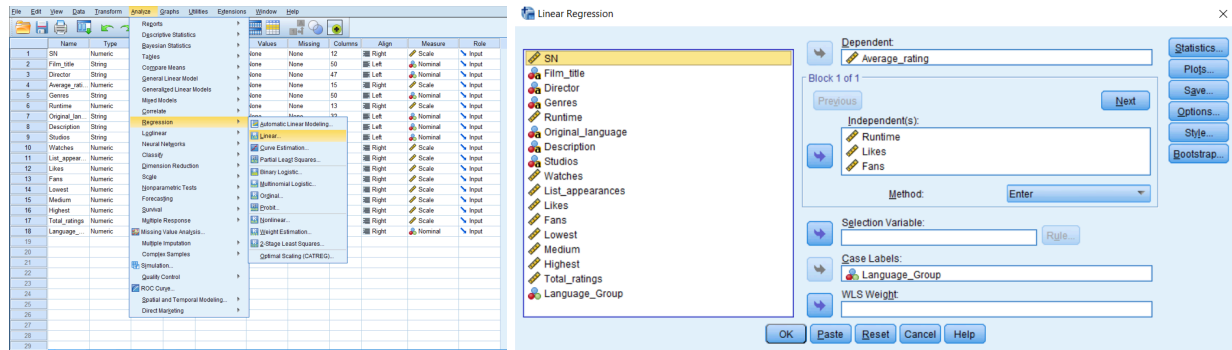


Figure 20: Regression Formula tab

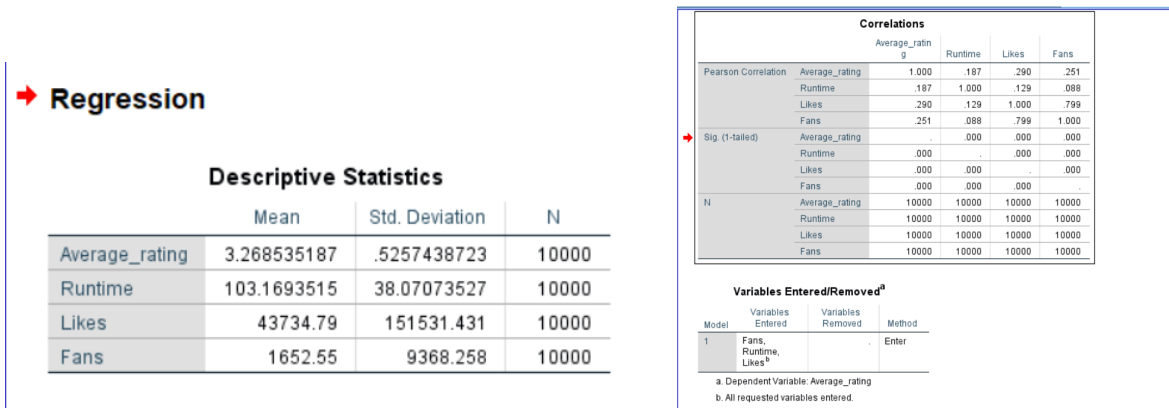


Figure 21: Regression statistics

Model Summary^b

Change Statistics

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.329 ^a	.108	.108	.4966252884	.108	403.305	3	9996	.000

a. Predictors: (Constant), Fans, Runtime, Likes

b. Dependent Variable: Average_rating

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	298.410	3	99.470	403.305	.000 ^b
	Residual	2465.380	9996	.247		
	Total	2763.790	9999			

a. Dependent Variable: Average_rating

b. Predictors: (Constant), Fans, Runtime, Likes

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Zero-order	Correlations		Collinearity Statistics	
		B	Std. Error				Lower Bound	Upper Bound		Partial	Part	Tolerance	VIF
1	(Constant)	3.012	.014		209.893	.000	2.984	3.040					
	Runtime	.002	.000	.153	16.005	.000	.002	.002	.187	.158	.151	.983	1.018
	Likes	7.764E-7	.000	.224	14.160	.000	.000	.000	.290	.140	.134	.357	2.799
	Fans	3.272E-6	.000	.058	3.706	.000	.000	.000	.251	.037	.035	.361	2.773

a. Dependent Variable: Average_rating

Figure 22: Regression Model Summary

Coefficient Correlations ^a					
Model		Fans	Runtime	Likes	
1	Correlations	Fans	1.000	.026	-.798
		Runtime	.026	1.000	-.099
		Likes	-.798	-.099	1.000
	Covariances	Fans	7.794E-13	3.031E-12	-3.862E-14
		Runtime	3.031E-12	1.732E-8	-7.119E-13
		Likes	-3.862E-14	-7.119E-13	3.006E-15
a. Dependent Variable: Average_rating					

Collinearity Diagnostics ^a							
Model	Dimension	Eigenvalue	Condition Index	Variance Proportions			
				(Constant)	Runtime	Likes	Fans
1	1	2.353	1.000	.01	.01	.03	.03
	2	1.397	1.298	.02	.02	.06	.08
	3	.188	3.534	.00	.00	.90	.89
	4	.061	6.190	.96	.97	.01	.00
a. Dependent Variable: Average_rating							

Residuals Statistics ^a					
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	3.016851187	6.311018944	3.268535187	.1727539886	10000
Residual	-2.18332458	1.536527634	.0000000000	.4965507815	10000
Std. Predicted Value	-1.457	17.612	.000	1.000	10000
Std. Residual	-4.396	3.094	.000	1.000	10000
a. Dependent Variable: Average_rating					

Figure 23: Coefficient for Regression Analysis

Charts

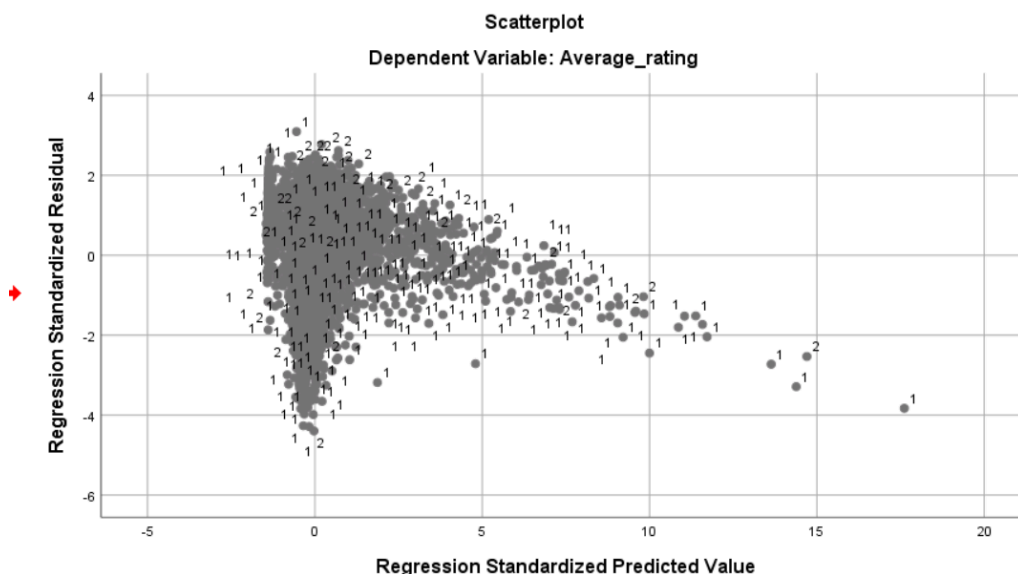


Figure 24: Scatter Plot for Regression model

The regression analysis also studied whether runtime, likes and fans predict the average rating of a film. The model has a statistically significant value ($F = 403.30$, p less than .001) yet it only predicted 10.8% of the variance ($R^2 = .108$) which is not that powerful. All the three predictors had considerable impact on ratings, likes had the strongest positive impact, then the runtime, and then fans. This implies that movies that get a higher number of likes or those with a greater duration of running time are rated slightly higher (Laerd.com, 2018).

The statistical analyses are very enlightening in the way language, runtime and engagement values are associated with the rating and the popularity of the movies in the dataset. The correlation analysis revealed that runtime is positively and significantly related to both watches and likes, which implies longer films receive a little bit more interest. The regression analysis showed that likes, runtime, and fans have a significant predictive effect on average rating, but the model has explained only 10.8% of the variance. It means that engagement is one of the factors that lead to ratings, but there are other factors like content quality, genre, or preference of the audience, which is much more significant in the process of rating a film.

Insights and Recommendations

The discussion has brought some notable conclusions regarding the behaviour of the audience, performance of the film, and the pattern of its engagement. There are always higher average ratings of the non-English films than English-language films even though the numbers are dominated by English titles. This implies that foreign or multi-cultural material can be more appealing to viewers. The engagement statistics (watches, likes, fans) possess a very high skew, with a very low percentage of movies receiving a significant amount of attention. Runtime has a poor yet positive correlation with engagement, which suggests that marginally longer movies can work better. The genre patterns display that the categories relating to drama have higher cumulative ratings because of the larger number of representation and that the niche genres are under used.

- The high-performing genres (drama, crime, and character-driven films) should be encouraged to enhance overall engagement.
- Showcase more non-English movies in special sections because the ratings of these movies are higher on average.
- Increase subtitle and dubbing in more than one language to increase worldwide viewing and reach.
- Deliver personalised suggestions to users with the help of engagement metrics (likes, watches, fans).
- Marketing should be done on films that have increasing fan base since it is a good predictor of the future.
- Normalise the genre tagging to enhance precise data and analytical consistency.
- Track run time and recommend the best targets to the content creators.

There are however a few disadvantages: the metrics of engagement are very unequal, the category of genres is inaccurate, and the dataset does not include contextual variables, including marketing budgets or the year of release. Further studies may consider sentiment analysis of

descriptions, time patterns or predictive models to include other variables to improve the accuracy of decision-making.

References

Baylor.edu. (2025). *Guides: SPSS Statistical Software: Graphs, Charts, and Visuals*. [online] Available at: <https://libguides.baylor.edu/c.php?g=1351162&p=10436062>

Excel-easy.com. (2025). *Descriptive Statistics in Excel*. [online] Available at: <https://www.excel-easy.com/examples/descriptive-statistics.html>.

Kent.edu. (2018). *LibGuides: SPSS Tutorials: Descriptive Stats for Many Numeric Variables (Descriptives)*. [online] Available at: <https://libguides.library.kent.edu/SPSS/Descriptives>.

Laerd.com. (2018). *Linear Regression Analysis in SPSS Statistics - Procedure, assumptions and reporting the output*. [online] Available at: <https://statistics.laerd.com/spss-tutorials/linear-regression-using-spss-statistics.php>.

Laerd.com. (2018). *Pearson's Product-Moment Correlation in SPSS Statistics - Procedure, assumptions, and output using a relevant example*. [online] Available at: <https://statistics.laerd.com/spss-tutorials/pearsons-product-moment-correlation-using-spss-statistics.php>.

Optimize Smart. (2025). *Optimize Smart*. [online] Available at: <https://www.optimizesmart.com/how-to-select-best-excel-charts-for-your-data-analysis-reporting/>.