TRAINITY PROJECT 5

MAYANK KUMAR GANDHARV

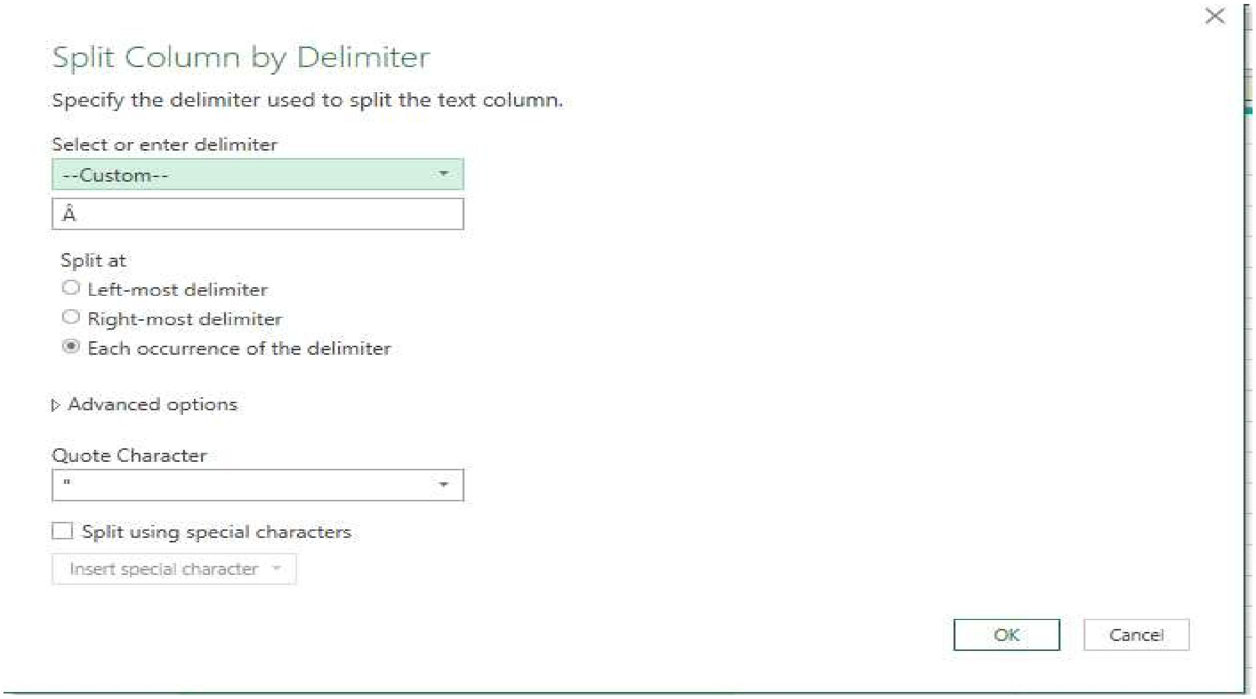
IMDB MOVIE ANALYSIS

HYPERLINK TO EXCEL FILE :
https://docs.google.com/spreadsheets/d/1wfODISUmSKhW1a2sS69dIeCThsWvYTM1/edit?usp=drive_link&ouid=106990321423670318865&rtpof=true&sd=true

Problem Statement: The dataset provided is related to IMDB Movies. A potential problem to investigate could be: "What factors influence the success of a movie on IMDB?" Here, success can be defined by high IMDB ratings. The impact of this problem is significant for movie producers, directors, and investors who want to understand what makes a movie successful to make informed decisions in their future projects.

TECH STACK USED: For this Project I used Microsoft Power BI and Microsoft Excel 2021.

DATA CLEANING: To clean data I used power query in Power BI as well Excel and table format in Excel to eliminate Duplicate Data. There are various columns that plays no role in the analysis but still I kept them in the model except column "Color". The movie_title has letter "Â" at the end of each title row so, in order to eliminate it Split Column by Delimiter is been used (shown in screenshot)



Split Column by Delimiter

Specify the delimiter used to split the text column.

Select or enter delimiter

--Custom--

Â

Split at

○ Left-most delimiter

○ Right-most delimiter

● Each occurrence of the delimiter

▷ Advanced options

Quote Character

"

□ Split using special characters

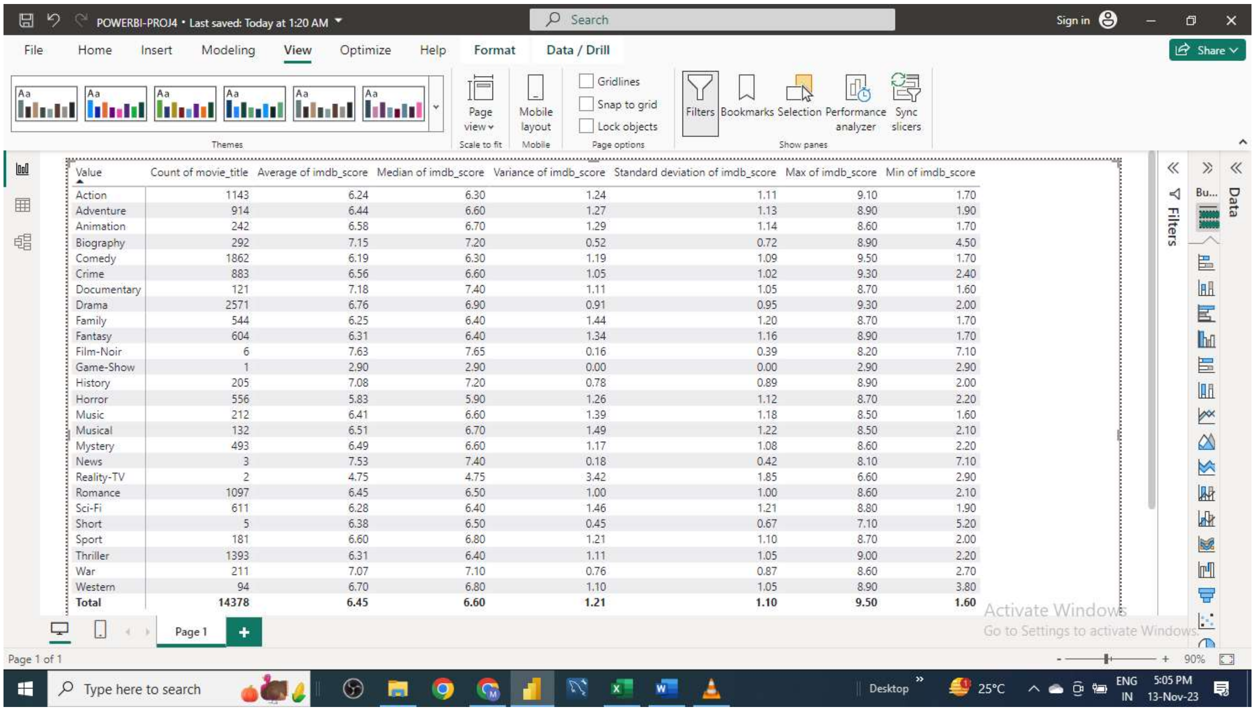
A. **Movie Genre Analysis:** Analyze the distribution of movie genres and their impact on the IMDB score.

For this task I used power BI mostly because in power BI, I was able to calculate the median of every genre based on the IMDB ratings. To calculate the Range and Mode I used Excel.

To perform Statistical functions I used following steps:

1. Loading the Data (CSV) into Power Query.
2. Splitting the data Using Delimiter "|" so that the genre could be distributed for each movie title
3. Lastly I Unpivoted the data so that I could get every genre for every Movie title and performed further steps by forming a matrix.
4. After loading the data into the worksheet I used delete duplicate function and deleted all the duplicates(256 duplicates)

The descriptive statistics for each genre (Count, Average, Median, Variance, Standard Deviation, Max and Min):



Value	Count of movie_title	Average of imdb_score	Median of imdb_score	Variance of imdb_score	Standard deviation of imdb_score	Max of imdb_score	Min of imdb_score
Action	1143	6.24	6.30	1.24	1.11	9.10	1.70
Adventure	914	6.44	6.60	1.27	1.13	8.90	1.90
Animation	242	6.58	6.70	1.29	1.14	8.60	1.70
Biography	292	7.15	7.20	0.52	0.72	8.90	4.50
Comedy	1862	6.19	6.30	1.19	1.09	9.50	1.70
Crime	883	6.56	6.60	1.05	1.02	9.30	2.40
Documentary	121	7.18	7.40	1.11	1.05	8.70	1.60
Drama	2571	6.76	6.90	0.91	0.95	9.30	2.00
Family	544	6.25	6.40	1.44	1.20	8.70	1.70
Fantasy	604	6.31	6.40	1.34	1.16	8.90	1.70
Film-Noir	6	7.63	7.65	0.16	0.39	8.20	7.10
Game-Show	1	2.90	2.90	0.00	0.00	2.90	2.90
History	205	7.08	7.20	0.78	0.89	8.90	2.00
Horror	556	5.83	5.90	1.26	1.12	8.70	2.20
Music	212	6.41	6.60	1.39	1.18	8.50	1.60
Musical	132	6.51	6.70	1.49	1.22	8.50	2.10
Mystery	493	6.49	6.60	1.17	1.08	8.60	2.20
News	3	7.53	7.40	0.18	0.42	8.10	7.10
Reality-TV	2	4.75	4.75	3.42	1.85	6.60	2.90
Romance	1097	6.45	6.50	1.00	1.00	8.60	2.10
Sci-Fi	611	6.28	6.40	1.46	1.21	8.80	1.90
Short	5	6.38	6.50	0.45	0.67	7.10	5.20
Sport	181	6.60	6.80	1.21	1.10	8.70	2.00
Thriller	1393	6.31	6.40	1.11	1.05	9.00	2.20
War	211	7.07	7.10	0.76	0.87	8.60	2.70
Western	94	6.70	6.80	1.10	1.05	8.90	3.80
Total	**14378**	**6.45**	**6.60**	**1.21**	**1.10**	**9.50**	**1.60**

STEPS TO CALCULATE MODE:

There are 2 approach to calculate mode for every genre. Since Mode is the most occurring value in a dataset,

Approach 1:

1. First filter the data to desired genre (lets say Action).
2. Then copy the data and paste it on another sheet.
3. Now, use **MODE.SNGL(Array)** Function to calculate the mode of the data.

Approach 2:

1. Using Pivot Table place the Genre and Imdb into "Rows" and then in "Values" place Count of Imdb
2. Now order the whole data column into descending order.
3. The first imdb that is shown under every genre has the most count and thus is the value of mode.

CALCULATING RANGE:

Select Max and Min in Pivot and subtract Min from Max for Each Genre.

Calculating Rest of the Descriptive Statistics (Average, Median, Var, StdDev)

Used Power BI to calculate these Descriptive Statistics using a Matrix

OUTPUT:

Genre	Count	Max	Min	Average	Variance	StdDev	Median	Range	mode
Action	1113	9.1	1.7	6.231626	1.252668	1.119226495	6.3	7.4	6.6
Adventure	888	8.9	1.9	6.436712	1.291458	1.136423313	6.6	7	6.7
Animation	240	8.6	1.7	6.575	1.309414	1.144296389	6.7	6.9	6.7
Biography	291	8.9	4.5	7.148797	0.525197	0.724704646	7.2	4.4	7
Comedy	1848	9.5	1.7	6.192857	1.190875	1.091272091	6.3	7.8	6.7
Crime	869	9.3	2.4	6.563061	1.059037	1.029095313	6.6	6.9	6.6
Documentary	121	8.7	1.6	7.180165	1.11627	1.056536782	7.4	7.1	7.5
Drama	2536	9.3	2	6.7653	0.909415	0.953632394	6.9	7.3	7.2
Family	534	8.7	1.7	6.23764	1.46419	1.210037311	6.4	7	6.7
Fantasy	583	8.9	1.7	6.302744	1.362054	1.167070825	6.4	7.2	6.7
Film-Noir	6	8.2	7.1	7.633333	0.186667	0.43204938	7.65	1.1	#N/A
Game-Show	1	2.9	2.9	2.9	#DIV/0!	#DIV/0!	2.9	0	#VALUE!
History	203	8.9	2	7.085714	0.786775	0.887003442	7.2	6.9	7.5
Horror	540	8.7	2.2	5.80463	1.255285	1.120394863	5.9	6.5	6.2
Music	212	8.5	1.6	6.406132	1.39982	1.183139907	6.6	6.9	6.5
Musical	131	8.5	2.1	6.500763	1.507769	1.227912311	6.7	6.4	7
Mystery	485	8.6	2.2	6.483918	1.174121	1.083568634	6.6	6.4	6.6
News	3	8.1	7.1	7.533333	0.263333	0.513160144	7.4	1	#N/A
Reality-TV	2	6.6	2.9	4.75	6.845	2.61629509	4.75	3.7	#N/A
Romance	1083	8.6	2.1	6.446076	0.997182	0.998589959	6.5	6.5	6.5

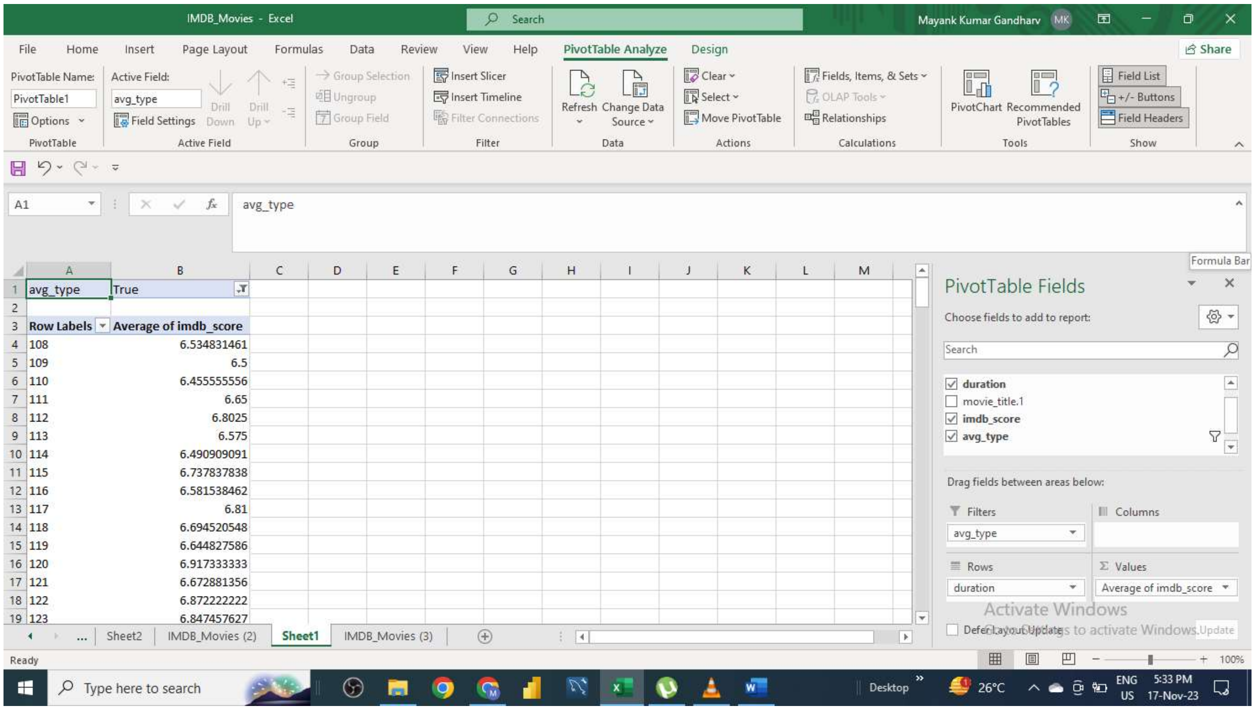
Sci-Fi	594	8.8	1.9	6.277778	1.480349	1.216695735	6.4	6.9	6.7
Short	5	7.1	5.2	6.38	0.557	0.746324326	6.5	1.9	#N/A
Sport	177	8.7	2	6.60113	1.230453	1.109257978	6.8	6.7	7.2
Thriller	1361	9	2.2	6.309111	1.116652	1.056717665	6.4	6.8	6.4
War	210	8.6	2.7	7.070952	0.767238	0.875921414	7.1	5.9	7.1
Western	94	8.9	3.8	6.703191	1.114506	1.055701584	6.8	5.1	6.8
Grand Total	**14130**	**9.5**	**1.6**	**6.447898**	**1.214286**	**1.101946321**	**6.6**	**7.9**	**6.7**

INSIGHT:

1. “Drama” genre is the most occurring movie genre in the dataset provided (2571 movies), followed by Comedy, Thriller, Action, Romance being the top 5 movie genre.
2. The most average IMDB Rating is given to the genre “Film-Noir” i.e 7.63 but there are very less movies for this genre. So for top 5 genre selection “Drama” is again the highest average IMDB Rated (6.76).
3. The highest IMDB rating is given to the movie: Towering Inferno (9.5) and the genre is “Comedy”.

B. **Movie Duration Analysis:** Analyze the distribution of movie durations and its impact on the IMDB score.
This task is done on Microsoft Excel 2021. To do this task I followed following steps:

1. First I loaded the data into my Excel’s Power Query and and cleaned the data by removing unwanted columns and applied the changes into the excel sheet
2. Then I deleted the duplicate data by selecting all the data (Ctrl+A) and removed duplicate rows by remove duplicates under “data”.
3. After removing duplicates i calculated the average of imdb score (so that i could segregate the data into 2 category, “above average” and “below average”).
4. After calculating average of duration, I selected the data and added one more column name “avg_type” and performed an “IF” Function so that I could category all the individual duration into above average by “TRUE” and below average by “FALSE” using syntax: **=IF((B2)> F3,"True","False")** where **F3** is the average value of duration (fixed).
5. After segregating the data based on average i created a Pivot Table.



	A	B
1	avg_type	True
2		
3	Row Labels	Average of imdb_score
4	108	6.534831461
5	109	6.5
6	110	6.455555556
7	111	6.65
8	112	6.8025
9	113	6.575
10	114	6.490909091
11	115	6.737837838
12	116	6.581538462
13	117	6.81
14	118	6.694520548
15	119	6.644827586
16	120	6.917333333
17	121	6.672881356
18	122	6.872222222
19	123	6.847457627

PivotTable Fields

Choose fields to add to report:

☑ duration
☐ movie_title.1
☑ imdb_score
☑ avg_type

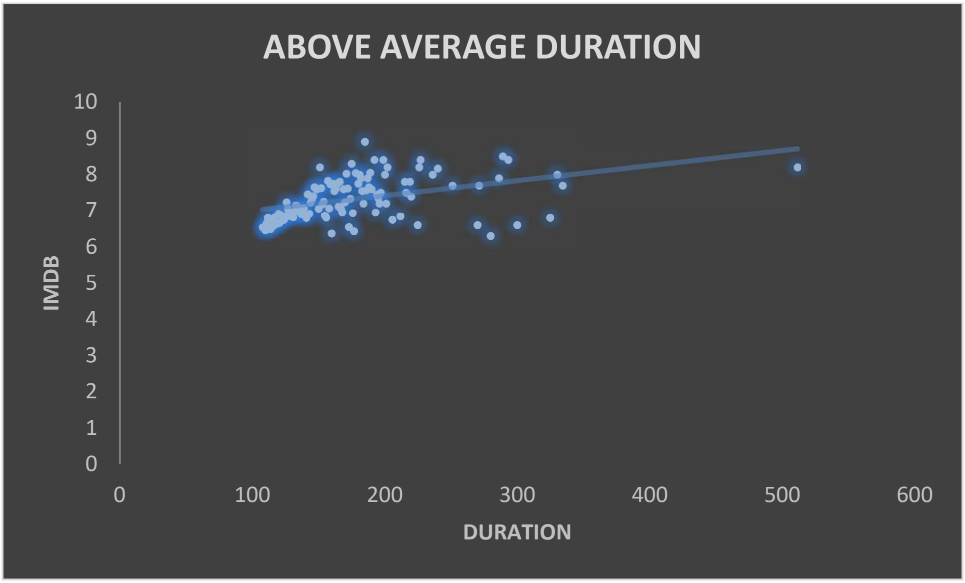
Drag fields between areas below:

Filters: avg_type
Columns:
Rows: duration
Values: Average of imdb_score

☐ Defer Layout Update

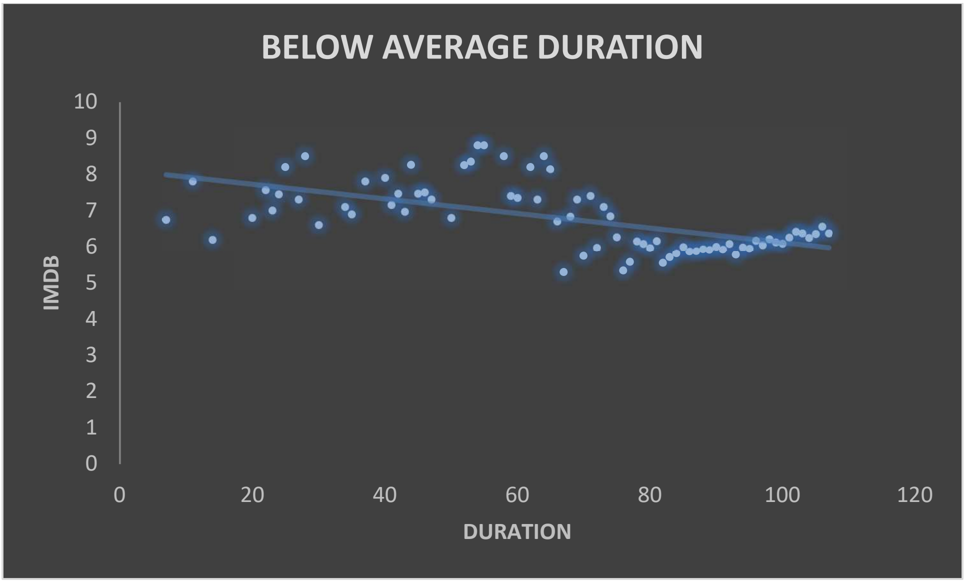
Here I calculated the average of imdb score of every distinct duration for example:
There are 99 movies with duration 110. So its easy to calculate the average of those movies imdb ratings rather than calculating imdb for each and every movie.

The scatter plot for movies that is above average duration:



OUTPUT:
For every movie whose duration is more than 107 have an increasing TRENDLINE with **Correlation Coefficient** of **0.43163** (=CORREL(ARRAY1, ARRAY2))

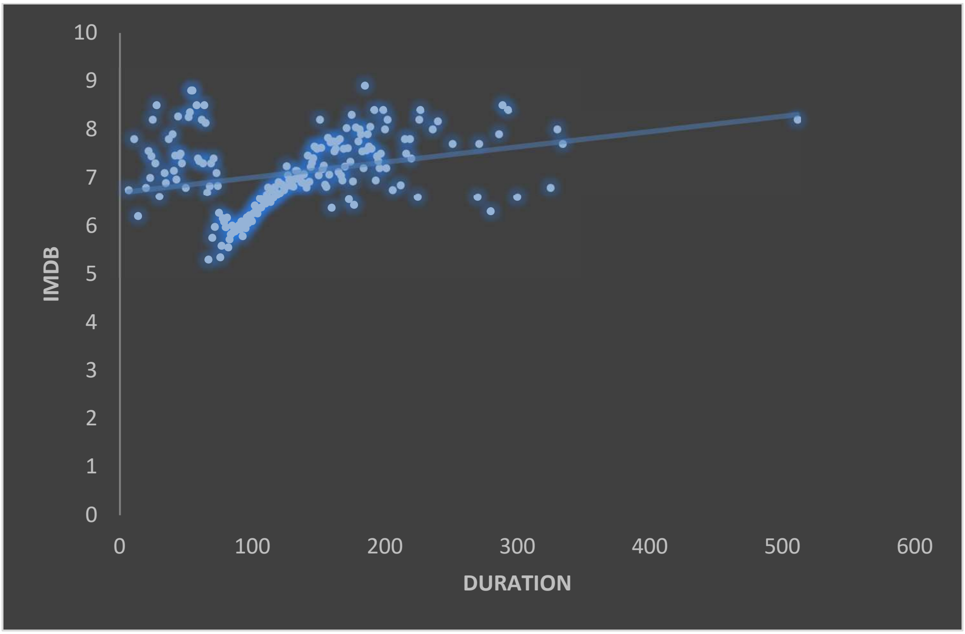
The scatter plot for movies that is below average duration :



OUTPUT:

For every movie whose duration is <= 107 have an decreasing TRENDLINE with **Correlation Coefficient** of **-0.58848** (=CORREL(ARRAY1, ARRAY2)).

For overall every movie duration in one scatter plot:



OUTPUT:

Here the trend is positive but first there is a sudden drop and then the graph is increasing (can be analyzed from the scatter plot). The correlation is 0.3.

INSIGHT:

It can be seen that movies with duration above 107 follows a positive trendline i.e with increasing movie duration the imdb also increases. But, movies below or equal to 107 have a negative trendline i.e with increasing movie duration the imdb decreases.

C. **Language Analysis:** Situation: Examine the distribution of movies based on their language.

For this task I followed following steps:

1. Using power query I removed columns that were irrelevant for the analysis (kept only Title_name, Language, IMDB Score.).
2. After loading the data from power query into the worksheet I then deleted the duplicate (**78** Duplicate data).
3. Lastly I did the analysis using Pivot Table. (for calculating count of movies and average of their IMDBs.)
4. To calculate the median for most occurring language I first sorted data in power query for that language and then loaded the whole sorted data into the **POWER BI DESKTOP** and calculated median by forming matrix.

Output of the pivot table:

language	Count of movie_title	Average	StdDev	Median
Aboriginal	2	6.95	0.777817459	6.95
Arabic	5	7.38	0.884307639	7.4
Aramaic	1	7.1	#DIV/0!	7.1
Bosnian	1	4.3	#DIV/0!	4.3
Cantonese	11	6.954545455	0.704788814	7.2
Chinese	3	5.666666667	0.550757055	5.7
Czech	1	7.4	#DIV/0!	7.4
Danish	5	7.5	1.077032961	8.1
Dari	2	7.5	0.141421356	7.5
Dutch	4	7.425	0.434932945	7.45
Dzongkha	1	7.5	#DIV/0!	7.5
English	4585	6.393740458	1.125155637	6.5
Filipino	1	6.7	#DIV/0!	6.7
French	73	7.038356164	0.726985812	7.2
German	19	7.342105263	0.954123093	7.6
Greek	1	7.3	#DIV/0!	7.3
Hebrew	5	7.58	0.334664011	7.6
Hindi	28	6.632142857	1.398955582	6.95
Hungarian	1	7.1	#DIV/0!	7.1
Icelandic	2	7.55	0.919238816	7.55
Indonesian	2	7.9	0.424264069	7.9
Italian	11	7.227272727	1.244259546	7.3

Japanese	17	7.347058824	1.000073527	7.5
Kannada	1	7.1	#DIV/0!	7.1
Kazakh	1	6	#DIV/0!	6
Korean	8	7.3875	0.825378701	7.5
Mandarin	24	6.7875	1.036848276	7.05
Maya	1	7.8	#DIV/0!	7.8
Mongolian	1	7.3	#DIV/0!	7.3
None	2	7.95	0.777817459	7.95
Norwegian	4	7.15	0.574456265	7.3
Panjabi	1	6.6	#DIV/0!	6.6
Persian	4	7.575	1.203813385	7.95
Polish	3	7.966666667	0.981495458	7.4
Portuguese	8	7.4875	0.883883476	7.7
Romanian	2	7.2	0.989949494	7.2
Russian	11	6.363636364	1.383671007	6.5
Slovenian	1	6.4	#DIV/0!	6.4
Spanish	40	6.9375	0.855056603	7.15
Swahili	1	7.4	#DIV/0!	7.4
Swedish	5	7.44	0.756967635	7.6
Tamil	1	5.1	#DIV/0!	5.1
Telugu	1	8.4	#DIV/0!	8.4
Thai	3	6.633333333	0.450924975	6.6
Urdu	1	7	#DIV/0!	7
Vietnamese	1	7.4	#DIV/0!	7.4
Zulu	2	7.1	0.282842712	7.1
Grand Total	**4908**	**6.436776691**	**1.127141831**	**7.3**

There is #DIV/0! Error in Standard Deviation because Standard Deviation of single value i.e 1 cant be calculated.

TO CALCULATE MEDIAN OF ENGLISH LANGUAGE:

Syntax used: **=MEDIAN(IMDB_Movies__4[imdb_score])**

OUTPUT OF MEDIAN:

6.5

INSIGHTS:

“ENGLISH” is the most used language for movies. The average of the IMDB for English Language is 6.40 (rounded), with Standard deviation 1.0125 and median IMDB for all the English Language is 6.5 (Medium Value). The most occurring IMDB value (Mode) for English Language is 6.7 (206 count).

D. **Director Analysis:** Influence of directors on movie ratings.
In order to do this task I followed following steps:

1. Select IMDB score, Director Name and Movie Title.
2. Deleting the duplicate
3. Select all the data (Ctr+A) and use pivot table for further analysis.
4. Selecting Director Name and average of IMDB.
5. Excluding the blank values from Director Name and sorting the Average of IMDB to Desc order.
6. To calculate the percentile I used **Percent Rank** function and also used **Percentile.INC** function.

OUTPUT:

Top 10 Director list:

Director Name	**Average of imdb_score**	Percentile using percent rank	imdb value based on percentile
John Blanchard	9.5	1	9.5
Sadyk Sher-Niyaz	8.7	0.998	8.6
Mitchell Altieri	8.7	0.998	8.6
Cary Bell	8.7	0.998	8.6
Mike Mayhall	8.6	0.997	8.5
Charles Chaplin	8.6	0.997	8.5
Raja Menon	8.5	0.996	8.4856
Ron Fricke	8.5	0.996	8.4856
Majid Majidi	8.5	0.996	8.4856
Damien Chazelle	8.5	0.996	8.4856

Syntax for Percent Rank: **=PERCENTRANK.INC($E$4:$E$2398,E4)**

Syntax for Percentile: **=PERCENTILE.INC($E$4:$E$2398,F4)**

E. **Budget Analysis:** Explore the relationship between movie budgets and their financial success.

For this task I followed following steps:

1. Sorted columns such as Gross, Budget and Movie Title from power query.
2. Did subtraction of gross to budget and made another column for it, named it Profit Margin.
3. Added the data to workbook and sorted the Profit Margin in Desc Order. Like that I got all the high profit movies of all time.
4. Lastly I found the correlation coefficient of Gross and Budget.

TOP-10 MOVIES ACCORDING TO PROFIT MARGIN ARE:

Movie Title	Profit Margin
Avatar	523505847
Jurassic World	502177271
Titanic	458672302
Star Wars: Episode IV - A New Hope	449935665
E.T. the Extra-Terrestrial	424449459
The Avengers	403279547
The Lion King	377783777
Star Wars: Episode I - The Phantom Menace	359544677
The Dark Knight	348316061
The Hunger Games	329999255

The Correlation Coefficient between Gross and Budget is 0.101033478

(Syntax: **=CORREL(IMDB_Movies__5[gross], IMDB_Movies__5[budget])**

Insights:

The movie that did the most profit is "Avatar" (523505847). The linear relation between the Gross Income and the Budget is somewhat straight (upward Direction) because the coefficient is 0.101.

Since Data is already been selected in descending order there is no need to use the MAX function in order to find the high profit making movies.