

# IMDB Movie Analysis

## Final Project-1

### Project Description:

The dataset provided is related to IMDB Movies.

Here, success can be defined by high IMDB ratings.

This impact 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.

### Approach:

We start by downloading the provided dataset and loading it on excel.

During the Data Cleaning process, we first delete the columns we dont require.

Next, we check for any blanks or discrepancies in the data and take care of it.

We delete duplicate records from certain columns.

We insert the data into a table.

Further we create various pivot tables according to our requirements.

We make use of Excel functions to derive insights from the data and show them graphically using charts.

### Tech-Stack Used:

#### Microsoft 360

Version : 18.2411.1163.0

Correlation ID : dac98643-f7f0-449f-aebb-fd9c99011baf

Session ID : 13fff215-657d-4760-ae33-ccab8bee0d16

# Excel

## Insights:

### Loading the dataset in excel file:

We start by opening Microsoft 360, then click on Excel.

When the window opens, we click on Upload → Locate our file → Press Okay.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	director_name	num_critic_for_reviews	duration	gross	genres	G2	G3	G4	G5	G6	G7	G8	language	budget	title_year
1	James Cameron	723	178	\$760,505,847.00	Action	Adventure	Fantasy	Sci-Fi					English	237000000	2
2	Gore Verbinski	302	169	\$309,404,152.00	Action	Adventure	Fantasy						English	300000000	2
3	Sam Mendes	602	148	\$200,074,175.00	Action	Adventure	Thriller						English	245000000	2
4	Christopher Nolan	813	164	\$448,130,642.00	Action	Thriller							English	250000000	2
5	Andrew Stanton	462	132	\$73,058,679.00	Action	Adventure	Sci-Fi						English	263700000	2
6	Sam Raimi	392	156	\$336,530,303.00	Action	Adventure	Romance						English	258000000	2
7	Nathan Greno	324	100	\$200,807,262.00	Adventure	Animation	Comedy	Family	Fantasy	Musical	Romance		English	260000000	2
8	Joss Whedon	635	141	\$458,991,599.00	Action	Adventure	Sci-Fi						English	250000000	2
9	David Yates	375	153	\$301,956,980.00	Adventure	Family	Fantasy	Mystery					English	250000000	2
10	Zack Snyder	673	183	\$330,249,062.00	Action	Adventure	Sci-Fi						English	250000000	2
11	Bryan Singer	434	169	\$200,069,408.00	Action	Adventure	Sci-Fi						English	209000000	2
12	Marc Forster	403	106	\$168,368,427.00	Action	Adventure							English	200000000	2
13	Gore Verbinski	313	151	\$423,032,628.00	Action	Adventure	Fantasy						English	225000000	2
14	Gore Verbinski	450	150	\$89,289,910.00	Action	Adventure	Western						English	215000000	2
15	Zack Snyder	733	143	\$291,021,565.00	Action	Adventure	Fantasy	Sci-Fi					English	225000000	2
16	Andrew Adamson	258	150	\$141,614,023.00	Action	Adventure	Family	Fantasy					English	225000000	2
17	Joss Whedon	703	173	\$623,279,547.00	Action	Adventure	Sci-Fi						English	220000000	2
18	Rob Marshall	448	136	\$241,063,875.00	Action	Adventure	Fantasy						English	250000000	2
19	Barry Sonnenfeld	451	106	\$179,020,854.00	Action	Adventure	Comedy	Family	Fantasy	Sci-Fi			English	225000000	2
20	Peter Jackson	422	164	\$255,108,370.00	Adventure	Fantasy							English	250000000	2
21	Marc Webb	599	153	\$262,030,663.00	Action	Adventure	Fantasy						English	230000000	2
22	Ridley Scott	343	156	\$105,219,735.00	Action	Adventure	Drama	History					English	200000000	2
23	Peter Jackson	509	186	\$258,355,354.00	Adventure	Fantasy							English	225000000	2
24	Chris Weitz	251	113	\$70,083,519.00	Adventure	Family	Fantasy						English	180000000	2
25	Peter Jackson	446	201	\$218,051,260.00	Action	Adventure	Drama	Romance					English	207000000	2
26	James Cameron	315	194	\$658,672,302.00	Drama	Romance							English	200000000	1
27	Anthony Russo	516	147	\$407,197,282.00	Action	Adventure	Sci-Fi						English	250000000	2
28	Peter Berg	377	131	\$65,173,160.00	Action	Adventure	Sci-Fi	Thriller					English	209000000	2
29	Colin Trevorrow	644	124	\$652,177,271.00	Action	Adventure	Sci-Fi	Thriller					English	150000000	2
30	Sam Mendes	750	143	\$304,360,377.00	Action	Adventure	Thriller						English	300000000	2

### Deleting Unwanted Columns:

We delete columns like Color, Facebook likes, Movie link, Actor Name, etc.

In the end, we're left with columns: Director Name, Duration, Gross, Genre, Language, Budget, imdb score.

These are the important columns that we will use for analysis.

### Handling Missing Data:

Select a column header and apply Sort & Filter function to filter out all blank values.

Apply this technique to each column.

You can either delete the rows with missing values or replace those cells with MEAN or MODE values of the column.

## Inserting a Table:

Select the complete data, then go to the insert tab, then click on the table option.

Your table has been created.

You can rename and customize your table according to your desire.

Dataset Table: [IMDB\\_Movies.xlsx](#)

## Pivot Table:

A PivotTable is an interactive way to quickly summarize large amounts of data. You can use a PivotTable to analyze numerical data in detail, and answer unanticipated questions about your data. A PivotTable is especially designed for: Querying large amounts of data in many user-friendly ways.

To make a Pivot table, select your table or columns that you need, go to Insert, click on PivotTable.

You get the option to create the Pivot table on a new spreadsheet or existing sheet. In that case you have to mention the location on the sheet.

Once the Pivot table has been created, you can rename and customize it according to requirement.

## Charts:

An Excel chart or graph is a visual representation of a Microsoft Excel worksheet's data. These graphs and charts allow you to see trends, make comparisons, pinpoint patterns, and glean insights from within the raw numbers. Excel includes countless options for charts and graphs, including bar, line, and pie charts.

### **To create a chart**

Select the data for which you want to create a chart.

Click INSERT > Recommended Charts.

On the Recommended Charts tab, scroll through the list of charts that Excel recommends for your data, and click any chart to see how your data will look. ...

When you find the chart you like, click it > OK.

## Excel Functions:

Excel has many premade formulas, called functions.

Functions are typed by = and the functions name.

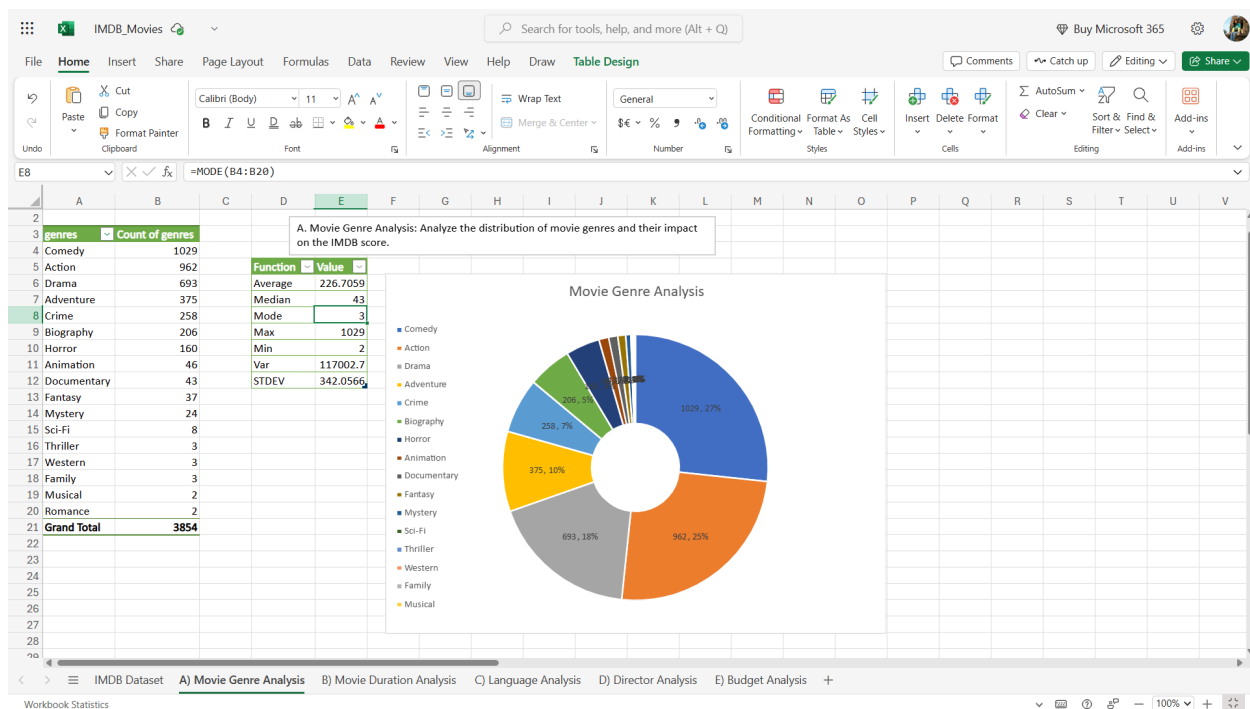
Once you have typed the function name you need to apply it to a range.

The range is always inside of parentheses.

## A)Movie Genre Analysis:

Analyze the distribution of movie genres and their impact on the IMDB score.

**Your Task:**Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores.



As the dataset contains multiple genres for each movie entry in a single column, we select Genre column then click on Split Text To Column option, in the Data ribbon option.

Next we create a pivot table, and enter Genres in row, and Count of Genres in  $\Sigma$  Values.

Using this Pivot Table, we insert a Pie Chart to visualize distribution of various movie genres.

Lastly, we use Excel functions `=AVERAGE(B4:B20)` to find average, `=MEDIAN(B4:B20)` to find median, `=MODE(B4:B20)` to find mode value, `=MAX(B4:B20)` to find maximum value,

=MIN(B4:B20)to find the minimum value, =VAR(B4:B20) for finding variance and =STDEV(B4:B20)to find Standard Deviation.

## Result:

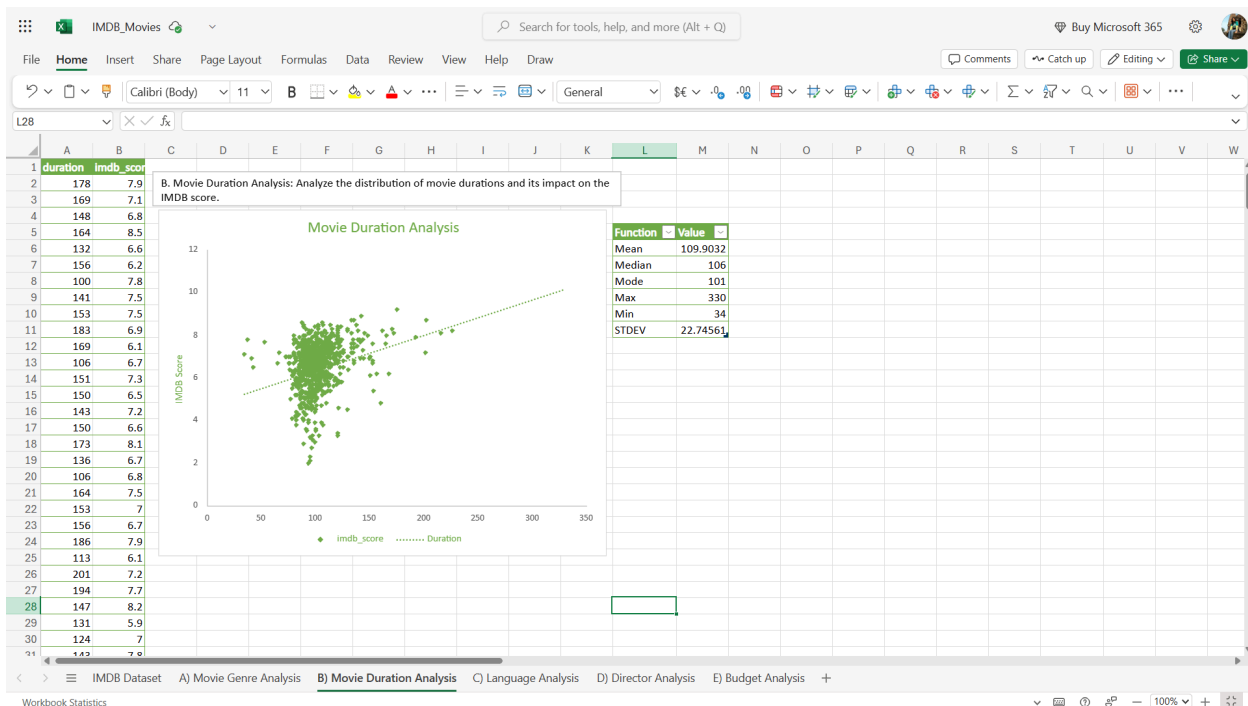
Most movies had been in the comedy genre, which is 1029 movies, that's 27% of total distribution and lowest is 2 in Musical and Romance genre.

Average count of genres is 226.7, Median is 43, Mode is 3, Variance is 117002.7 and Standard Deviation is 342.

## B) Movie Duration Analysis:

Analyze the distribution of movie durations and its impact on the IMDB score.

**Your Task:** Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.



First we copy paste Duration and imdb\_score columns in another sheet, then we insert them in a table.

Then we use this table to create a chart. We are using a column chart over here.

We also find values of Mean, Median, Mode, Min, Max and Stdev for movie duration.

## Result:

The chart shows that the majority of movies are around 100 minutes and mostly have around 6. The linear trendline goes from 30 minutes, 5 score to 340 minutes, 8 score.

Values of Mean, Median, Mode, Min, Max and Stdev for movie duration are 109.9, 106, 101, 330, 34 and 22.74 respectively.

## C) Language Analysis:

Situation: Examine the distribution of movies based on their language.

**Your Task:** Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.

Language	Aboriginal	Cantonese	Chinese	Dutch	English	French	German	Hindi	Italian	Japanese	Korean	Mandarin	Portuguese	Russian	Spanish	Swedish	Thai
Count	2	9	3	3	3653	53	11	17	5	13	7	22	2	10	20	3	3
Average	6.3	6.1875	5.9	6.8	5.539189	6.570968	6.9125	6.36154	6.94	6.07692308	6.242857	6.636842105	7.55	7.0444444	6.76	6.4666667	6.9
Median	6.3	6.25	6.7	7.6	5.55	6.5	6.7	6.5	6.9	6.3	6.5	6.7	7.55	6.9	6.7	6.7	6.7
STDEV	0.565685425	0.642400632	1.7435596	1.47309	2.177017	1.349616	0.7754031	1.41217	0.541295	1.43767975	0.927105	0.814596373	0.353553391	0.6326751	0.711939	0.6806859	0.3

We create a pivot table and put language and imdb\_score to rows. Then we put  $\Sigma$  Values to Count of Language.

We use Excel functions to find the Average, Median and Standard Deviation distribution of imdb\_score for each language.

Then we insert it in a table.

## Result:

The table mentions Count, Average, Median and STDEV for languages having more than one count.

Lan	Cant				Chi D				Portu								
gua	Abori	ones	ne	ut	Engli	Frenc	Germ		Italia	Japa	Kore	Man	gues	Russi	Spani	Swed	
ge	ginal	e	se	ch	sh	h	an	Hindi	n	nese	an	darin	e	an	sh	ish	Thai
Cou																	
nt	2	9	3	3	3653	53	11	17	5	13	7	22	2	10	20	3	3
Ave					6.906	6.703	5.628	6.083					6.183	6.686			
rag					4.8493	3333	5714	3333					7.22	3333	6666		
e	1.75	3.15	3.88	15	33	29	33	6.25	5	33	67	7.55	6	6.625	5.95	7.1	
Me																	
dia					4.												
n	1.75	3.15	3.88	6.9	6.7	6	6.1	6.25	7.35	6.5	6.8	7.55	6.1	6.5	5.95	7.1	
	0.212	0.244			1.308	0.665	1.446	1.590	0.129	0.64	0.847	1.030	0.353	1.24	0.536	0.494	0.707
STD	1320	9489	0.		3416	1332	5064	7878	0994	5438	1520	8572	5533	4989	5231	9747	1067
EV	34	74	0.11	95	94	65	74	45	54	92	48	91	96	99	47	81	

## D) Director Analysis:

Influence of directors on movie ratings.

**Your Task:** Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

director_name	Average of Imdb_score	Percentile Rank
Charles Chaplin	8.6	0.998
Tony Kaye	8.6	0.998
Alfred Hitchcock	8.5	0.996
Damien Chazelle	8.5	0.996
Ron Fricke	8.5	0.996
Majid Majidi	8.5	0.996
Sergio Leone	8.433333333	0.996
Christopher Nolan	8.425	0.995
Asghar Farhadi	8.4	0.993
Marius A. Markevicius	8.4	0.993
Richard Marquand	8.4	0.993
S.S. Rajamouli	8.4	0.993
Billy Wilder	8.3	0.99
Lee Unkrich	8.3	0.99
Fritz Lang	8.3	0.99
Lenny Abrahamson	8.3	0.99
Pete Docter	8.233333333	0.99
Hayao Miyazaki	8.225	0.989
Quentin Tarantino	8.2	0.989
George Roy Hill	8.2	0.986
Elia Kazan	8.2	0.986
Joshua Oppenheimer	8.2	0.986
Juan Jos�� Campanella	8.2	0.986
Victor Fleming	8.15	0.986
Milos Forman	8.133333333	0.985
David Singleton	8.1	0.981
Tony Gatlif	8.1	0.981

We insert a pivot table, put director\_name in rows and Average of imdb\_score in  $\Sigma$  Values.

We use `=PERCENTRANK.EXC($B$4:$B$1755, B4)` to find percent rank for each row.

## E) Budget Analysis:

Explore the relationship between movie budgets and their financial success.

**Your Task:** Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

	gross	budget	profit
1	\$760,505,847.00	\$237,000,000.00	\$523,505,847.00
2	\$309,404,152.00	\$300,000,000.00	\$9,404,152.00
3	\$200,074,175.00	\$245,000,000.00	-\$44,925,825.00
4	\$448,130,642.00	\$250,000,000.00	\$198,130,642.00
5	\$73,058,679.00	\$263,700,000.00	-\$190,641,321.00
6	\$336,530,303.00	\$258,000,000.00	\$78,530,303.00
7	\$200,807,262.00	\$260,000,000.00	-\$59,192,738.00
8	\$458,991,599.00	\$250,000,000.00	\$208,991,599.00
9	\$301,956,980.00	\$250,000,000.00	\$51,956,980.00
10	\$330,249,062.00	\$250,000,000.00	\$80,249,062.00
11	\$200,069,408.00	\$209,000,000.00	-\$8,930,592.00
12	\$168,368,427.00	\$200,000,000.00	-\$31,631,573.00
13	\$423,032,628.00	\$225,000,000.00	\$198,032,628.00
14	\$89,289,910.00	\$215,000,000.00	-\$125,710,090.00
15	\$291,021,565.00	\$225,000,000.00	\$66,021,565.00
16	\$141,614,023.00	\$225,000,000.00	-\$83,385,977.00
17	\$623,279,547.00	\$220,000,000.00	\$403,279,547.00
18	\$241,063,875.00	\$250,000,000.00	-\$8,936,125.00
19	\$179,020,854.00	\$225,000,000.00	-\$45,979,146.00
20	\$255,108,370.00	\$250,000,000.00	\$5,108,370.00
21	\$262,030,663.00	\$230,000,000.00	\$32,030,663.00
22	\$105,219,735.00	\$200,000,000.00	-\$94,780,265.00
23	\$258,355,354.00	\$225,000,000.00	\$33,355,354.00
24	\$70,083,519.00	\$180,000,000.00	-\$109,916,481.00
25	\$218,051,260.00	\$207,000,000.00	\$11,051,260.00
26	\$658,672,302.00	\$200,000,000.00	\$458,672,302.00
27	\$407,197,282.00	\$250,000,000.00	\$157,197,282.00
28	\$65,173,160.00	\$209,000,000.00	-\$143,826,840.00
29	\$652,177,271.00	\$150,000,000.00	\$502,177,271.00
30	\$264,360,777.00	\$200,000,000.00	\$64,360,777.00

We copy-paste the columns gross and budget columns in a new sheet.

Then we add a third column using the mathematical function '-' subtract to find profit by subtracting gross from budget.

After selecting all three columns we put them in a table.

The **CORREL** function returns the **correlation coefficient** of two cell **ranges**. Use the correlation coefficient to determine the relationship between two properties.

We use excel function `=CORREL (A:A, B:B)` to find the correlation coefficient of the table.

We use the excel function `=MAX (C2:C3855)` to find the amount of highest profit.

## Result:

The value of Correlation Coefficient is 0.10.



The value of movie having maximum profit is  
\$523,505,847.00

Hiring Process Analytics Workbook: [IMDB\\_Movies.xlsx](#)