



MOVIE ANALYSIS

- 1) **PROJECT DESCRIPTION:** “What factors influence the success of a movie on IMBD?” This project aims to answer that question using various parameters like duration, genre, director etc.

This project involves working with data to analyze it and derive insights from it that contribute towards better and profitable filmmaking for producers.



- 2) **APPROACH: 1st STEP: Data Cleaning –**

This step involves sorting and reprocessing data to make it suitable for analysis. The dataset used was cleaned in the following ways-

- 1) Removed columns, empty and unwanted values to prepare the data for analysis.
- 2) Cleaned values with unwanted characters for proper readability.

2nd STEP: Data Analysis –

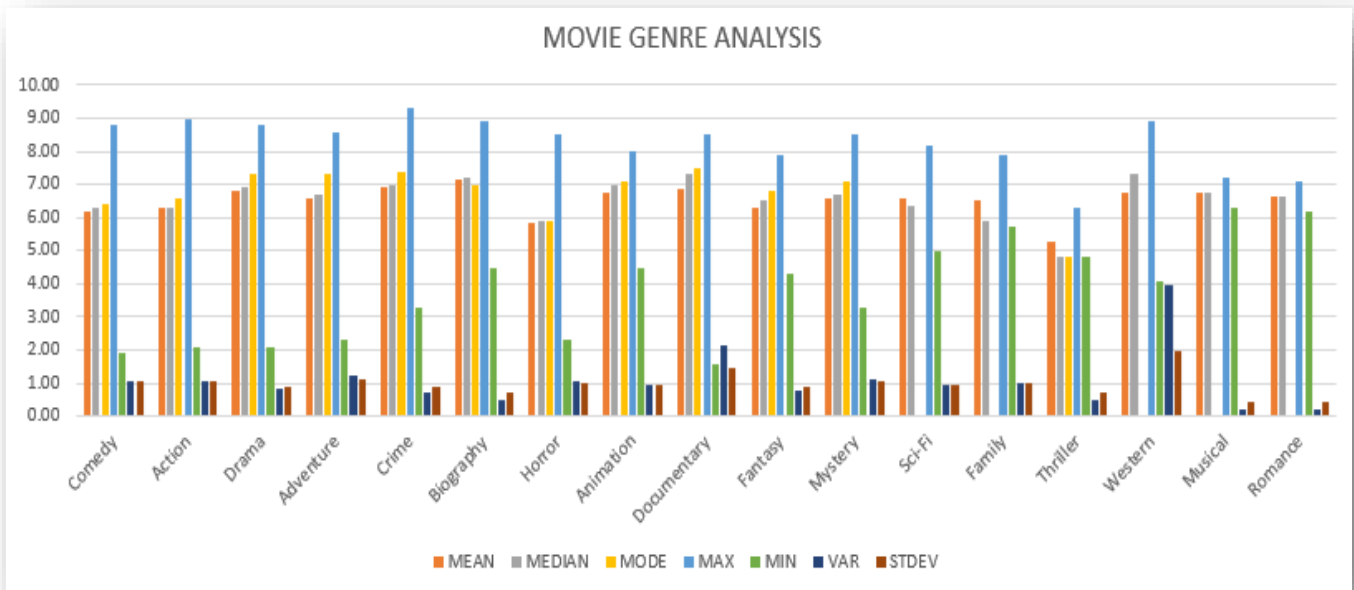
The cleaned, processed data is then analyzed and insights are taken from it using methods like root cause analysis. Various questions are asked till we get down to the root of a problem to find solutions for it. This expands our parameters and understanding and brings out our full potential.

- 3) **TECH-STACK USED:** MS Word  and MS Excel  were used to execute this project. MS Excel was used to clean, prepare, analyze and derive insights from data and MS Word was used to present this data.

4) INSIGHTS:

A. MOVIE GENRE ANALYSIS: Analyze the distribution of movie genres and their impact on the IMDB score.

TASK: DETERMINE THE MOST COMMON GENRES OF MOVIES IN THE DATASET. THEN, FOR EACH GENRE, CALCULATE THE DESCRIPTIVE STATISTICS (MEAN, MEDIAN, MODE, RANGE, VARIANCE, STANDARD DIVISION) OF THE IMDB SCORES.



GENRE	NUMBER OF MOVIES	MEAN	MEDIAN	MODE	MAX	MIN	VAR	STDEV
Comedy	1036	6.17	6.30	6.40	8.80	1.90	1.08	1.04
Action	969	6.29	6.30	6.60	9.00	2.10	1.08	1.04
Drama	699	6.81	6.90	7.30	8.80	2.10	0.82	0.91
Adventure	375	6.55	6.70	7.30	8.60	2.30	1.25	1.12
Crime	258	6.95	7.00	7.40	9.30	3.30	0.75	0.86
Biography	208	7.15	7.20	7.00	8.90	4.50	0.48	0.69
Horror	165	5.85	5.90	5.90	8.50	2.30	1.06	1.03
Animation	45	6.74	7.00	7.10	8.00	4.50	0.92	0.96
Documentary	38	6.88	7.30	7.50	8.50	1.60	2.16	1.47
Fantasy	37	6.28	6.50	6.80	7.90	4.30	0.78	0.88
Mystery	24	6.61	6.70	7.10	8.50	3.30	1.14	1.07
Sci-Fi	8	6.59	6.35	#N/A	8.20	5.00	0.93	0.96
Family	3	6.50	5.90	#N/A	7.90	5.70	0.99	0.99
Thriller	3	5.30	4.80	4.80	6.30	4.80	0.50	0.71
Western	3	6.77	7.30	#N/A	8.90	4.10	3.98	2.00
Musical	2	6.75	6.75	#N/A	7.20	6.30	0.20	0.45
Romance	2	6.65	6.65	#N/A	7.10	6.20	0.20	0.45

The most common genre of movies are **Comedy, Action, Drama, Adventure, Crime, Biography** and **Horror**. The **maximum number of movies** are made in the **Comedy** genre at **1036 movies** made in the genre.

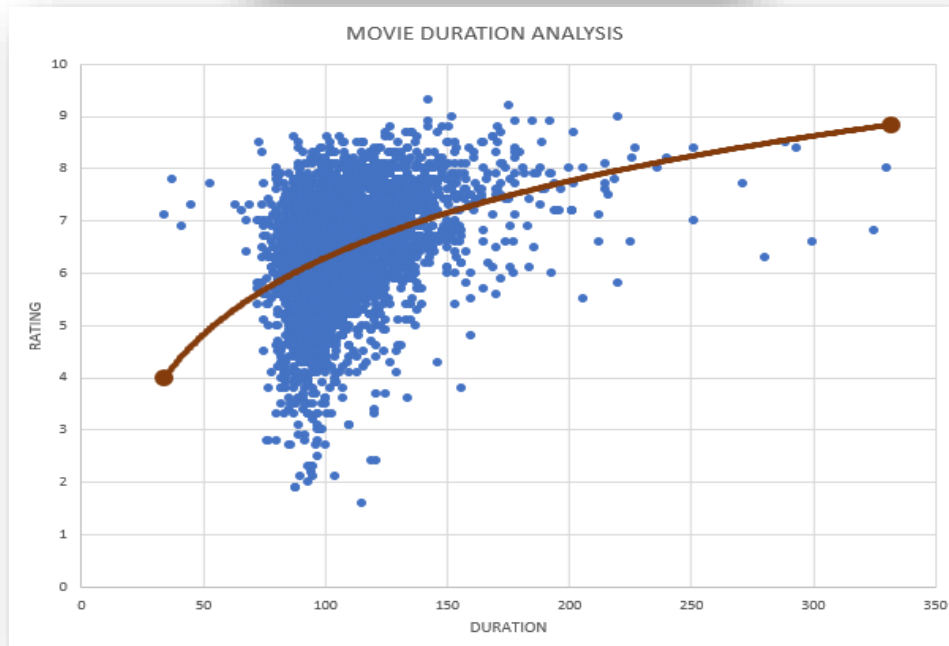
FORMULAS AND CONCEPTS APPLIED:

- ❖ Number of Movies: **Count** of the number of movies in each genre.
=COUNTIF(N:N,[@GENRE])
- ❖ Mean: The **average** rating of a particular genre.
=AVERAGEIF(N:N,[@GENRE], M:M)
- ❖ Median: The **mid value** of the rating of a genre.
=MEDIAN(IF(N:N=[@GENRE],M:M))
- ❖ Mode: The **most frequently occurring** rating of a particular genre. The genres marked N/A are the genres with no repeated ratings.
=MODE.SNGL(IF (N:N=[@GENRE], M:M))
- ❖ Max: The **maximum rating** that was given to a genre.
=MAXIFS (M:M, N:N,[@GENRE])
- ❖ Min: The **minimum rating** that was given to a genre.
=MINIFS (M:M, N:N,[@GENRE])
- ❖ Variance: Measure of how the ratings for a genre **spread out from the mean**.
=VAR.P(IF(N:N=[@GENRE],M:M))
- ❖ Standard Deviation: Quantifies the **amount of variation** between the ratings of the genre. It is the **square root of variance**.
=STDEV.P(IF (N: N=[@GENRE],M:M))

B. MOVIE DURATION ANALYSIS: Analyze the distribution of movie durations and its impact on the IMDB score.

TASK: ANALYZE THE DISTRIBUTION OF MOVIE DURATIONS AND IDENTIFY THE RELATIONSHIP BETWEEN MOVIE DURATION AND IMDB SCORE.

STATISTICAL FUNCTION	VALUE
MEAN	109.96
MEDIAN	106.00
MODE	101.00
STDEV	22.69



The **average** duration is **109 minutes**. The **maximum number of movies** have a duration of **101 minutes**.

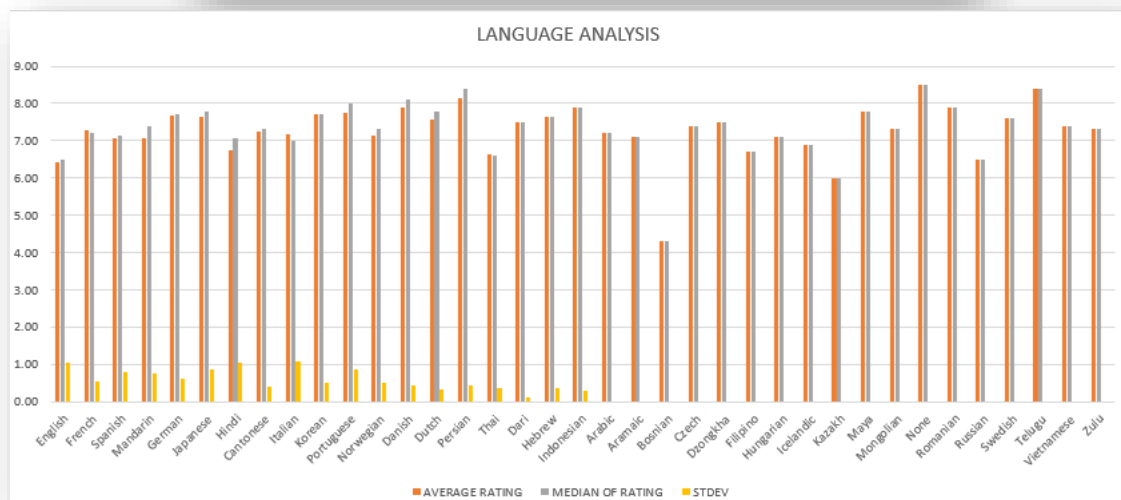
FORMULAS AND CONCEPTS APPLIED:

- ❖ **MEAN:** The **average** duration of movies.
=AVERAGE(C:C)
- ❖ **MEDIAN:** The **mid-point value** of the duration of movies.)
=MEDIAN(C:C)
- ❖ **MODE:** T The **most frequently occurring** duration of a movie.
=MODE.SNGL(C:C)
- ❖ **STDEV:** Measure of the amount of **variation between** different durations.
=STDEV.P(C:C)

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

TASK: DETERMINE THE MOST COMMON LANGUAGES USED IN MOVIES AND ANALYZE THEIR IMPACT ON THE IMDB SCORE USING DESCRIPTIVE ANALYSIS.

LANGUAGE	NUMBER OF MOVIES	AVERAGE RATING	MEDIAN OF RATING	STDEV
English	3697.00	6.42	6.50	1.05
French	37.00	7.29	7.20	0.55
Spanish	26.00	7.05	7.15	0.81
Mandarin	15.00	7.08	7.40	0.75
German	13.00	7.69	7.70	0.62
Japanese	12.00	7.63	7.80	0.86
Hindi	10.00	6.76	7.05	1.05
Cantonese	8.00	7.24	7.30	0.41
Italian	7.00	7.19	7.00	1.07
Korean	5.00	7.70	7.70	0.51
Portuguese	5.00	7.76	8.00	0.88
Norwegian	4.00	7.15	7.30	0.50
Danish	3.00	7.90	8.10	0.43
Dutch	3.00	7.57	7.80	0.33
Persian	3.00	8.13	8.40	0.45
Thai	3.00	6.63	6.60	0.37
Dari	2.00	7.50	7.50	0.10
Hebrew	2.00	7.65	7.65	0.35
Indonesian	2.00	7.90	7.90	0.30
Arabic	1.00	7.20	7.20	0.00
Aramaic	1.00	7.10	7.10	0.00
Bosnian	1.00	4.30	4.30	0.00
Czech	1.00	7.40	7.40	0.00
Dzongkha	1.00	7.50	7.50	0.00
Filipino	1.00	6.70	6.70	0.00
Hungarian	1.00	7.10	7.10	0.00
Icelandic	1.00	6.90	6.90	0.00
Kazakh	1.00	6.00	6.00	0.00
Maya	1.00	7.80	7.80	0.00
Mongolian	1.00	7.30	7.30	0.00
None	1.00	8.50	8.50	0.00
Romanian	1.00	7.90	7.90	0.00
Russian	1.00	6.50	6.50	0.00
Swedish	1.00	7.60	7.60	0.00
Telugu	1.00	8.40	8.40	0.00
Vietnamese	1.00	7.40	7.40	0.00
Zulu	1.00	7.30	7.30	0.00



The **highest number** of movies are made in **English** at **3697** movies. English is **followed by French** and **Spanish** at **37** and **26** movies respectively.

FORMULAS AND CONCEPTS APPLIED:

- ❖ NUMBER OF MOVIES: The number of movies in each language.
=COUNTIF(I:I,Table1!\$Q48)
- ❖ MEAN: The **average** of rating of movies in each language.
=AVERAGEIFS(M:M,I:I,Table1!\$Q48)
- ❖ MEDIAN: The **mid-point value** of rating of movies in each language.
=MEDIAN(IF(I:I=Table1!\$Q48,M:M))
- ❖ STDEV: Measure of the amount of **variation between** different rating of movies in each language.
=STDEV.P(IF(I:I=Table1!\$Q48,M:M))

D. DIRECTOR ANALYSIS: Influence of directors on IMDB ratings.


TASK: IDENTIFY THE TOP DIRECTORS BASED ON THEIR AVERAGE IMDB SCORE AND ANALYZE THEIR CONTRIBUTION TO THE SUCCESS OF MOVIES USING PERCENTILE CALCULATIONS.

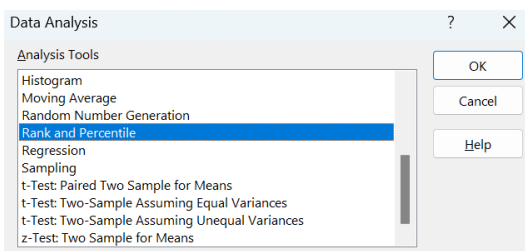
POINT	AVERAGE	PERCENTILE
Charles Chaplin	8.60	99.90%
Tony Kaye	8.60	99.90%
Alfred Hitchcock	8.50	99.70%
Damien Chazelle	8.50	99.70%
Majid Majidi	8.50	99.70%
Ron Fricke	8.50	99.70%
Sergio Leone	8.43	99.60%
Christopher Nolan	8.43	99.50%
Asghar Farhadi	8.40	99.30%
Marius A. Markevicius	8.40	99.30%
Richard Marquand	8.40	99.30%
S.S. Rajamouli	8.40	99.30%
Billy Wilder	8.30	99.10%
Fritz Lang	8.30	99.10%
Lee Unkrich	8.30	99.10%
Lenny Abrahamson	8.30	99.10%
Pete Docter	8.23	99.00%
Hayao Miyazaki	8.23	99.00%

The top directors are **Charlie Chaplin** and **Tony Kaye** with a shared **IMDB average of 8.60**. They come under the **top 1.1%** of the directors.

The **top 1%** directors are shown in the picture of the tables obtained.

FORMULAS AND CONCEPTS APPLIED:

- ❖ I used the  Data Analysis function found under the Analysis section on the Data Ribbon of excel.



- ❖ And used the Ranks and Percentiles function to obtain the necessary percentiles based on the average IMDB score.
- ❖ I converted that data into a table and presented it.

Note: The full table obtained is in the dataset provided after the Insights section.

E. BUDGET ANALYSIS: Explore the relationship between movie budgets and their financial success.

TASK: ANALYSE THE CORRELATION BETWEEN MOVIE BUDGETS AND GROSS EARNINGS, AND IDENTIFY THE MOVIES WITH THE HIGHEST PROFIT MARGIN.

movie name	gross	budget	profit margin
Avatar	760505847	237000000	523505847
Jurassic World	652177271	150000000	502177271
Titanic	658672302	200000000	458672302
Star Wars: Episode IV - A New Hope	460935665	110000000	449935665
E.T. the Extra-Terrestrial	434949459	105000000	424449459
The Avengers	623279547	220000000	403279547
The Lion King	422783777	450000000	377783777
Star Wars: Episode I - The Phantom Menace	474544677	115000000	359544677
The Dark Knight	533316061	185000000	348316061
The Hunger Games	407999255	78000000	329999255
Deadpool	363024263	58000000	305024263

CORRELATION COEFFICIENT	0.10
-------------------------	------

The movie with the **highest profit-margin** is **Avatar**, at a **profit margin** of **52,35,05,847**. The is **followed by Jurassic World, Titanic, Star Wars: Episode IV – A New Hope** and so on.

The **correlation coefficient** between **movie budgets** and **gross earnings** is a **0.10** indicating a **positive correlation** between movie budgets and gross earnings.

FORMULAS AND CONCEPTS APPLIED:

- ❖ **PROFIT MARGIN:** Gross earnings-Budget of the movies.
=([@gross]-[@budget])
- ❖ **CORRELATION COEFFICIENT:** A number between 1 and -1 indicating the strength and relationship between variables.
=CORREL(AU:AU,AT:AT)

LINK TO THE DATASET:

https://docs.google.com/spreadsheets/d/1ykCtQMLj-1NgJ_mk5pS2Yaf3atd2Gis4/edit?usp=sharing&oid=109740116170106135071&rtpof=true&sd=true

5) **RESULT:** Through this project, I achieved a detailed data-driven understanding of the various factors that influence the success of a movie on IMDB. I understood -

- popular genre of movies
- the language dynamics of movies
- how directors can have an effect on IMDB scores
- the correlation between budgets and gross earnings.

I used various formulas and functions to analyse data and derive insights from it. The valuable experience I got from this project has improved my skills as a Data Analyst.