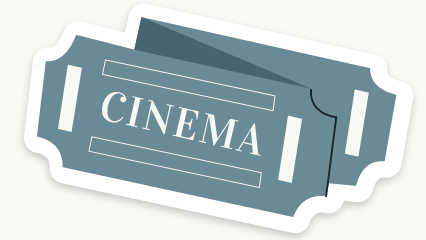


PROJECT 5

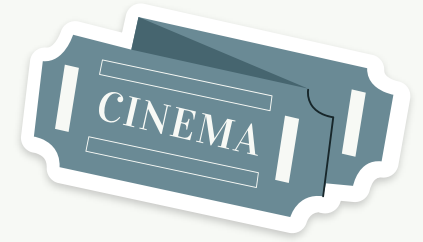
IMDB MOVIE ANALYSIS



WHAT IS IMDB?

IMDb stands for the Internet Movie Database. It is an online database that provides information about films, television programs, video games, and streaming content. IMDb includes details such as cast and crew credits, release dates, box office gross, trivia, reviews, and ratings submitted by users.





IMPORTANCE

1. Audience Influence:

Shapes viewing decisions through ratings and reviews.

2. Industry Impact:

Vital for market research, casting, and project planning.

3. Trend Analysis:

Unveils insights into genre preferences and audience trends.

4. Community Engagement:

Fosters a film-loving community through user discussions.

5. Research and Academia:

Essential resource for academic studies in film and culture.





PROJECT DESCRIPTION

Welcome to our presentation on IMDb movie analysis . In this project we will be using the knowledge of statistics and excel to do the tasks for analysis. Discover the importance of movie analysis and how it can provide valuable insights for filmmakers, movie enthusiasts, and decision-makers in the industry. It helps in understanding the factors influencing the success of a movie such as director, genre etc.





APPROACH

DATA CLEANING

Removing outliers, handling missing values , removing duplicates from the dataset

DATA ANALYSIS

Using stats methods and excel functions for summarizing the key metrics

DATA VISUALIZATION

Using charts and graphs for comprehensive data visualization





TECH STACK USED



MICROSOFT EXCEL:

- It is a versatile tool for collecting and organising data.
- It is used for data analysis including sorting , filtering and statistical calculations.
- It is used for creating visualizations (charts/graphs)

MICROSOFT POWERPOINT:

Final report is visualized in the form of presentation





DATA CLEANING



1. Removed duplicates
2. Removed rows with blanks which were important aspects of the movie like –
 - Duration of the movie
 - Director
 - Title year
 - Gross
 - Budget
 - Language
 - Country
 - Content rating
3. Titles of the movies had “Â” in them which had to be cleaned . So , replaced it with a blank space



A. Movie Genre Analysis

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.

1. Selecting the column genres and converting the genres into multiple columns via the text to columns option in the data option of the excel ribbon using the delimiter “|”
2. Making a list of unique genres
3. Using countif function , finding the count of each genre in movies

=COUNTIF(\$A:\$G,\$J2) where column a to g = genres converted to multiple columns and j2 is the first unique genre

Writing the above formula for the first genre and dragging it to apply for all the remaining genres

4. Using mean , median , mode , stdev, var, min , max functions on count of genres for descriptive statistics.

GENRES	COUNT
Action	954
Documentary	53
Adventure	781
Drama	1905
Animation	198
Comedy	1481
Mystery	381
Fantasy	509
Crime	707
Biography	239
Sci-Fi	495
Horror	388
Romance	864
Thriller	1107
Game-Show	0

Family	445
Music	156
Western	57
Musical	98
Film-Noir	1
History	149
Sport	150
War	153
News	0
Reality-TV	0
Short	0
TOTAL	11271

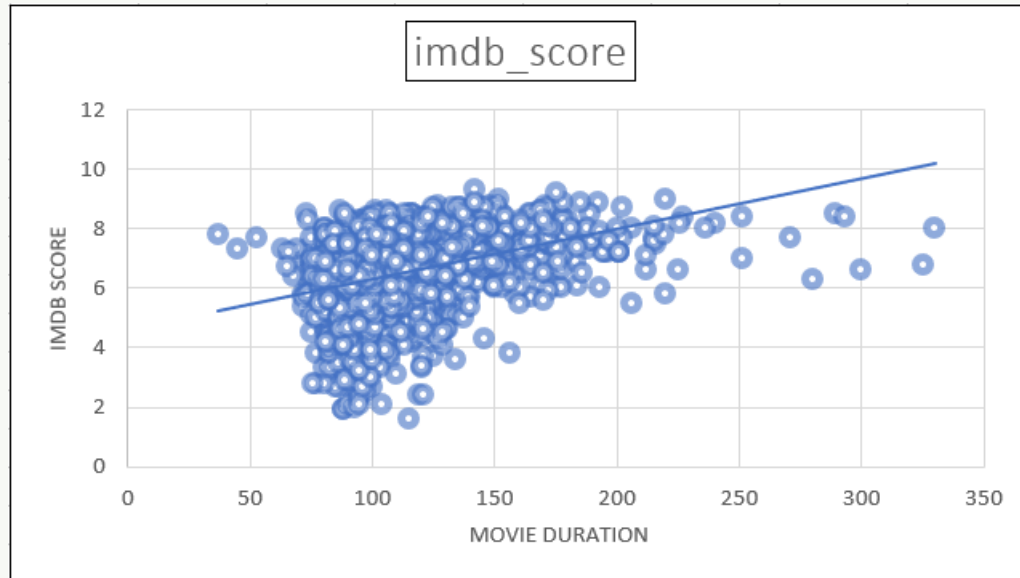
MEAN	433.5
MEDIAN	218.5
MODE	0
STDEV	485.5825
VAR	245221.9
MIN	0
MAX	1905

Drama and Comedy
are the two most
popular genres

B. Movie Duration Analysis

Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score

1. Selecting duration and imdb score columns.
2. Inserting a scatter plot.
3. Inserting a trendline and formatting it .
4. Using mean , median , stdev excel functions on movie durations for the descriptive statistics



MEAN	MEDIAN	STDEV
110.0571	106	22.6425

C. Language Analysis

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

1. Selecting the language column and inserting a pivot table
2. Inserting the count column in pivot table
3. Using averageif function to find the mean of each language

=AVERAGEIF(A:A,\$D2,B:B) where a:a is the column of language and b:b is the column of imdb score , d2 is the first language in the pivot table.

Writing the above formula for the first language and dragging it to apply for all the remaining languages

4. Finding mean and stdev using excel functions
5. Median and stdev will be same for all the languages

Language	Count of language	MEAN
Aboriginal	2	6.95
Arabic	1	7.2
Aramaic	1	7.1
Bosnian	1	4.3
Cantonese	7	7.34286
Czech	1	7.4
Danish	3	7.9
Dari	2	7.5
Dutch	3	7.56667
English	3625	6.42568
Filipino	1	6.7
French	34	7.35588
German	10	7.77
Hebrew	2	7.65
Hindi	5	7.22
Hungarian	1	7.1
Indonesian	2	7.9
Italian	7	7.18571

Japanese	10	7.66
Kazakh	1	6
Korean	5	7.7
Mandarin	14	7.02143
Maya	1	7.8
Mongolian	1	7.3
None	1	8.5
Norwegian	4	7.15
Persian	3	8.13333
Portuguese	5	7.76
Romanian	1	7.9
Russian	1	6.5
Spanish	23	7.08261
Thai	3	6.63333
Vietnamese	1	7.4
Zulu	1	7.3
(blank)		
Grand Total	3783	

MEDIAN	STDEV
6.6	1.05336

English is the most used language in movies

D. Director Analysis

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

1. Selecting the director name and imdb score columns
2. Inserting a pivot table
3. Changing the field settings of imdb score column which will be sum of imdb score as default **TO** average of imdb scores.
4. Sorting the directors name via the value filter – top 10 , sorting desc on the basis of average of imdb scores
5. Finding the large of the avg imdb scores via `=large(b2:b13,1)` WHERE B2:B13 ARE AVG IMDB SCORES
6. Finding the rank via `=PERCENTRANK(B2:B13,D1)` WHERE D1 IS LARGE
7. Finding the percentile via `=PERCENTILE(B2:B13,D2)` WHERE D2 IS RANK

Director name	Average of imdb_score
Akira Kurosawa	8.7
Tony Kaye	8.6
Charles Chaplin	8.6
Alfred Hitchcock	8.5
Ron Fricke	8.5
Majid Majidi	8.5
Damien Chazelle	8.5
Sergio Leone	8.433333333
Christopher Nolan	8.425
Marius A. Markevicius	8.4
Richard Marquand	8.4
Asghar Farhadi	8.4
Grand Total	8.466666667


large	8.7
rank	1
percentile	8.7

Akira Kurosawa is the most successful director with the highest average imdb score of 87

E. Budget Analysis


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

1. Using the budget and gross columns
2. Creating a new column named profit
3. Using the formula `=A2-B2` WHERE a2 is the gross earning for the first movie and b2 is the budget of the first movie
Dragging the formula for all the movies for finding the profit/loss of each movie
4. Using CORREL excel function to calculate the correlation coefficient between two sets of values
5. Finding highest profit margin using the max excel function.



profit
523505847
9404152
-44925825
198130642
-190641321
78530303
-59192738
208991599
51956980
80249062

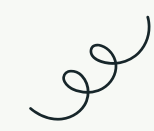
-8930592
-31631573
198032628
-125710090
66021565
-83385977
403279547
-8936125
-45979146
5108370
32030663
-94780265



And so on...

correlation	high profit margin
0.100005839	523505847

The movie with highest profit margin "523505847" is "AVATAR"
Directed by "James Cameron" in the year 2009





RESULT

- Most common genres
- Impact of genres on imdb score
- Relation between movie duration and imdb score
- Most common languages
- Impact of languages on imdb score
- Top 10 directors
- Impact of directors on imdb score
- Profit/ loss in each movie

