IMDB
MOVIE ANALYSIS

MRUNALI PETAKAR

PROJECT DESCRIPTION

This project aims to conduct a comprehensive investigation into the factors that contribute to a movie's success on IMDB, with a focus on high IMDB ratings. It is crucial for movie producers, directors, and investors to understand the elements that make a movie successful to make informed decisions for their future projects. The project involves meticulous data cleaning, thorough data analysis, utilizing the five 'Whys' approach, and presenting a detailed report and data story. The project's tasks involve analyzing movie genres, duration, language, directors, and budget to determine their impact on IMDB ratings. The project's findings will be useful for movie producers, directors, and investors to make informed decisions for their future projects.

APPROACH

my approach to this project was to use a structured approach to data analysis, starting with data cleaning and exploration, and moving on to more advanced techniques such as hypothesis testing and modeling. I used Excel's built-in functions and formulas to perform my analysis and create visualizations to communicate my findings.

TECH STACK USED

- Microsoft Excel: I am using Excel for data cleaning, analysis, and visualization.
- Google Drive: I am uploading the dataset and creating a shareable link for the project submission.

CLEANING THE DATA

This is one of the most important steps to perform before moving forward with the analysis. This step involves preprocessing the data to make it suitable for analysis. It includes handling missing values, removing duplicates, converting data types if necessary, and possibly feature engineering.

There were 5043 rows & 28 columns before cleaning. after cleaning 3724 rows & 10 columns are remaining.

TASK A:

Genre	Count of genres
Action	951
Adventure	773
Animation	196
Biography	238
Comedy	1455
Crime	704
Documentary	45
Drama	1876
Family	440
Fantasy	504
Film-Noir	1
History	147
Horror	386
Music	149
Musical	96
Mystery	378
Romance	851
Sci-Fi	492
Sport	147
Thriller	1105
War	150
Western	57
(blank)	
Grand Total	**11141**

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, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores.

I used Pivot tables, filter and sorting for this task

MEAN	506.4090909
MEDIAN	382
MODE	147
MAX	1876
MIN	1
VAR	243633.3009
SD	493.5922415

So from above 2 screenshots we can see that-

MOST COMMON / POPULOR GENRE - DRAMA

TASK A:

- descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores for each genre is found as below.

	AVERAGE	MEDIAN	MODE	RANGE	VARIANCE	SD
Comedy	6.861077	6.9	6.7	5.3	0.659369	0.812015
Thriller	9.2	9.2	9.2	0	0	0
Action	6.114301	6.2	6.7	7.4	1.303437	1.141682
Document	7.430769	7.6	7.6	4.9	1.146215	1.070614
Drama	6.496061	6.6	6.5	6.6	1.147997	1.071446
Crime	6.351779	6.5	7	6.9	1.126396	1.061318
Biography	6.621078	6.7	6.9	4.5	0.762953	0.873472
Family	7.375	7.5	8.9	3.3	2.5025	1.581929
Adventure	6.432425	6.6	6.7	6.2	1.221924	1.105407
Animation	6.864444	6.8	6.8	3.7	0.667343	0.816911
Mystery	5.656522	5.8	5.2	4.3	0.925296	0.961923
Horror	6.015723	6.1	6.1	5.7	0.766144	0.875296
Fantasy	6.143243	6.2	6.1	3.5	0.611967	0.782283
Musical	7.466667	7.6	8.7	2.8	1.973333	1.404754
Sci-Fi	6.175	6.5	8.7	4.7	2.836429	1.68417
Western	6.05	6.05	8.7	1.9	1.805	1.343503
Romance	7.3	7.3	8.7	3	4.5	2.12132

TASK B:

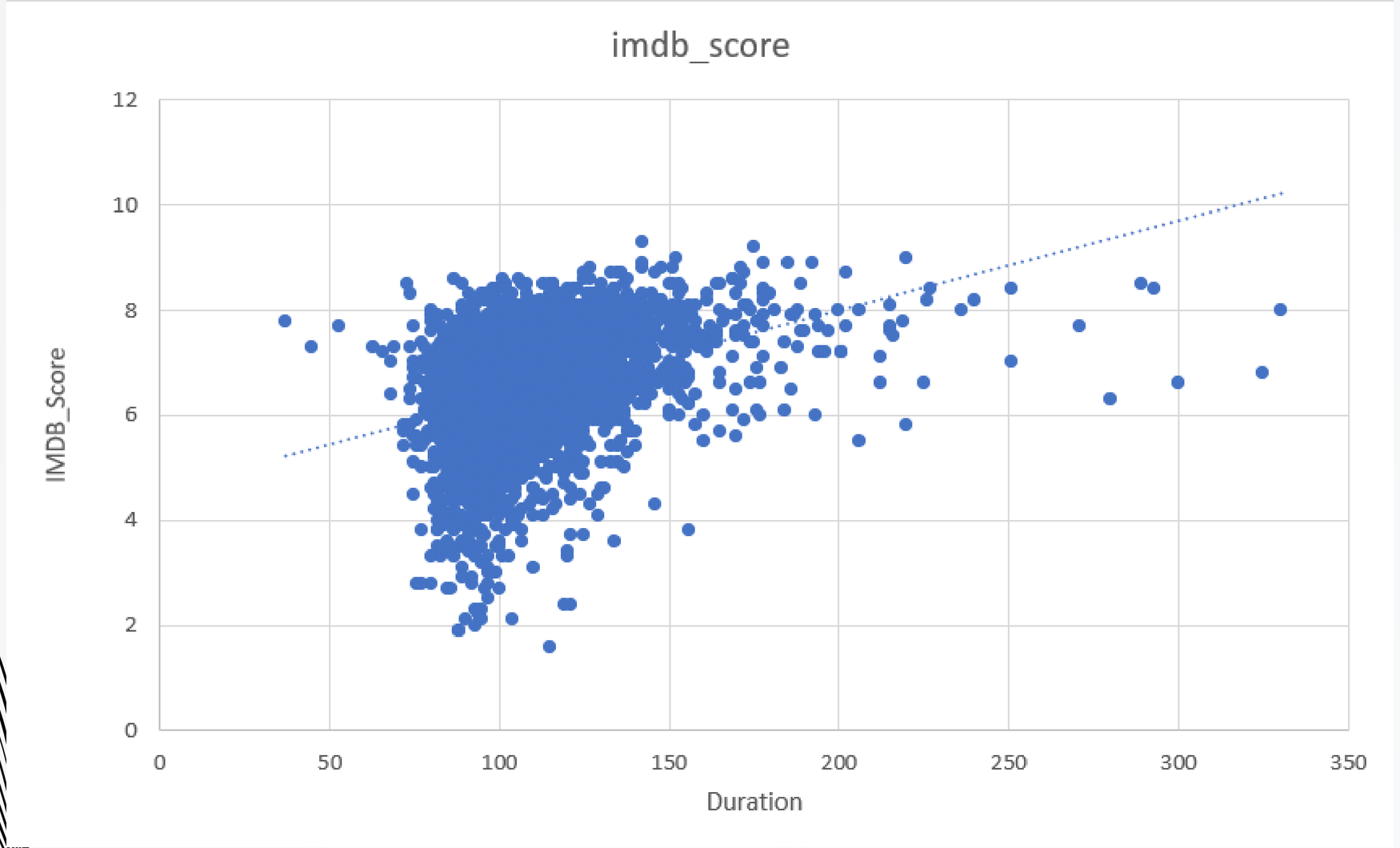
Movie Duration Analysis: Analyze the distribution of movie durations and its impact on the IMDB score.

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

1] Calculated descriptive statistics as below

mean	110.2635
median	106
sd	22.67832

2] scatter plot to visualize the relationship between movie duration and IMDB score as below



INSIGHTS:

INSIGHTS FROM TASK B (MOVIE DURATION ANALYSIS):

- THE DISTRIBUTION OF MOVIE DURATIONS IS SKEWED TO THE RIGHT, WITH A LONGER TAIL ON THE RIGHT SIDE.
- THERE IS A WEAK POSITIVE CORRELATION BETWEEN MOVIE DURATION AND IMDB SCORE, INDICATING THAT LONGER MOVIES TEND TO HAVE SLIGHTLY HIGHER RATINGS.
- THE MEAN MOVIE DURATION IS 110 MINUTES, THE MEDIAN DURATION IS 106 MINUTES, AND THE STANDARD DEVIATION IS 22.678 MINUTES.

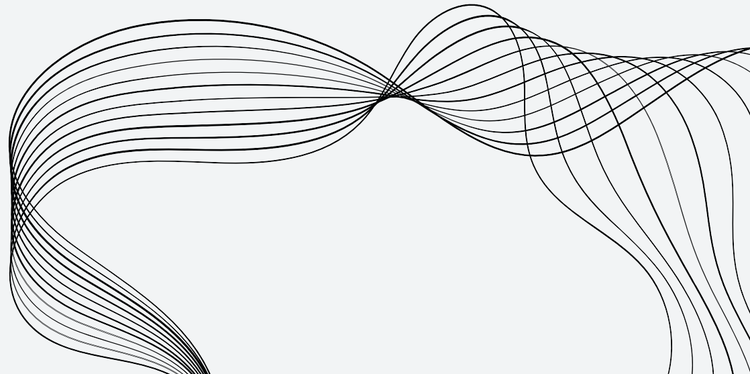
Row Labels	Count of language
Aboriginal	2
Arabic	1
Aramaic	1
Bosnian	1
Cantonese	7
Czech	1
Danish	3
Dari	2
Dutch	3
English	3566
Filipino	1
French	34
German	10
Hebrew	1
Hindi	5
Hungarian	1
Indonesian	2
Italian	7
Japanese	10
Kazakh	1
Korean	5
Mandarin	14
Maya	1
Mongolian	1
None	1
Norwegian	4
Persian	3
Portuguese	5
Romanian	1
Russian	1
Spanish	23
Thai	3
Vietnamese	1
Zulu	1
Grand Total	3723

TASK 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 statistics.

The adjacent table shows counts of languages and we can see the most common language is **ENGLISH** as the count is 3566



mean, median, and standard deviation of the IMDB scores for each language

Row Labels	Sum of imdb_score	Count of language	average_score	median	sd
Aboriginal	13.9	2	6.95	6.95	0.77782
Arabic	7.2	1	7.2	7.2	0
Aramaic	7.1	1	7.1	7.1	0
Bosnian	4.3	1	4.3	4.3	0
Cantonese	51.4	7	7.342857143	7.3	0.35051
Czech	7.4	1	7.4	7.4	0
Danish	23.7	3	7.9	8.1	0.52915
Dari	15	2	7.5	7.5	0.14142
Dutch	22.7	3	7.566666667	7.8	0.40415
English	22920.5	3566	6.428647737	6.5	1.0479
Filipino	6.7	1	6.7	6.7	0
French	250.1	34	7.355882353	7.3	0.51944
German	77.7	10	7.77	7.8	0.71188
Hebrew	8	1	8	8	0
Hindi	36.1	5	7.22	7.4	0.80125
Hungarian	7.1	1	7.1	7.1	0
Indonesian	15.8	2	7.9	7.9	0.42426
Italian	50.3	7	7.185714286	7	1.15532
Japanese	76.6	10	7.66	8	0.99017
Kazakh	6	1	6	6	0
Korean	38.5	5	7.7	7.7	0.57009
Mandarin	98.3	14	7.021428571	7.25	0.76579
Maya	7.8	1	7.8	7.8	0
Mongolian	7.3	1	7.3	7.3	0
None	8.5	1	8.5	8.5	0
Norwegian	28.6	4	7.15	7.3	0.57446
Persian	24.4	3	8.133333333	8.4	0.55076
Portuguese	38.8	5	7.76	8	0.97877
Romanian	7.9	1	7.9	7.9	0
Russian	6.5	1	6.5	6.5	0
Spanish	162.9	23	7.082608696	7.2	0.86058
Thai	19.9	3	6.633333333	6.6	0.45092
Vietnamese	7.4	1	7.4	7.4	0
Zulu	7.3	1	7.3	7.3	0
Grand Total	**24071.7**				

INSIGHTS:

INSIGHTS FROM TASK C (LANGUAGE ANALYSIS):

- THE MOST COMMON LANGUAGE USED IN MOVIES IS ENGLISH, FOLLOWED BY FRENCH AND SPANISH.
- THE MEAN IMDB SCORE FOR ENGLISH MOVIES IS 6.428, FOR SPANISH MOVIES IS 7.08, AND FOR FRENCH MOVIES IS 7.35.
- THE MEDIAN IMDB SCORE FOR ENGLISH MOVIES IS 6.5, FOR SPANISH MOVIES IS 7.2, AND FOR FRENCH MOVIES IS 7.3.
- THE STANDARD DEVIATION OF IMDB SCORES FOR ENGLISH MOVIES IS 1.0479, FOR SPANISH MOVIES IS 0.86, AND FOR FRENCH MOVIES IS 0.5194.

TASK D:

Influence of directors on movie ratings.
Task: Identify the top directors based on their average IMDB score

this is the list of top 20 directors based on average imdb score

directors	avg_by_director
Akira Kurosawa	8.7
Charles Chaplin	8.6
Tony Kaye	8.6
Damien Chazelle	8.5
Majid Majidi	8.5
Ron Fricke	8.5
Alfred Hitchcock	8.5
Sergio Leone	8.433333333
Christopher Nolan	8.425
Richard Marquand	8.4
Asghar Farhadi	8.4
Lenny Abrahamson	8.3
Lee Unkrich	8.3
Fritz Lang	8.3
Billy Wilder	8.3
Pete Docter	8.233333333
Hayao Miyazaki	8.225
Quentin Tarantino	8.2
Juan José Campanella	8.2
Joshua Oppenheimer	8.2

here we can see contribution to the success of movies using percentile calculations.
the percentrank column shows % contribution of director in movie success

directors	avg_by_direc	percentrank
Akira Kurosawa	8.7	1
Tony Kaye	8.6	1
Victor Fleming	8.15	1
William Wyler	8.1	1
Wolfgang Becker	7.7	1
Yimou Zhang	7.525	1
Zack Snyder	7.175	1
Zal Batmanglij	6.9	1
Charles Chaplin	8.6	0.999
Damien Chazelle	8.5	0.997
Majid Majidi	8.5	0.997
Alfred Hitchcock	8.5	0.996
Christopher Nolan	8.425	0.996
Ron Fricke	8.5	0.996
Sergio Leone	8.433333333	0.995
Asghar Farhadi	8.4	0.994
Fritz Lang	8.3	0.994
Lenny Abrahamson	8.3	0.993
Billy Wilder	8.3	0.992
Hayao Miyazaki	8.225	0.992

INSIGHTS:

INSIGHTS FROM TASK D (DIRECTOR ANALYSIS):
THE TOP DIRECTORS BASED ON THEIR AVERAGE IMDB SCORE ARE DIRECTOR AKIRA KUROSAVA, DIRECTOR TONY KAYE, AND DIRECTOR CHARLS CHAPLIN.

TASK E:

Explore the relationship between movie budgets and their financial success.

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

movie with highest profit margin	highest profit		correlation coeff
Avatar	523505847		0.098318102

the correlation coefficient between movie budgets and gross earnings using Excel's CORREL function will be 0.098318

and the movie with highest profit is AVATAR

INSIGHTS:

INSIGHTS FROM TASK E (BUDGET ANALYSIS):

- THERE IS A POSITIVE CORRELATION BETWEEN MOVIE BUDGETS AND GROSS EARNINGS, INDICATING THAT HIGHER BUDGETS TEND TO RESULT IN HIGHER EARNINGS.
- THE CORRELATION COEFFICIENT BETWEEN MOVIE BUDGETS AND GROSS EARNINGS IS 0.098318, INDICATING A WEAK POSITIVE RELATIONSHIP.
- THE MOVIE WITH THE HIGHEST PROFIT MARGIN IS MOVIE AVATAR, WITH A PROFIT MARGIN OF RS. 523505847.

I'M GIVING EXCEL FILE LINK HERE

RESULTS:

From this project, I have learned the importance of data cleaning and preprocessing to make the data suitable for analysis. This includes handling missing values, removing duplicates, and converting data types if necessary. Additionally, feature engineering can be used to create new variables that may be useful in the analysis.
I have also learned the importance of data analysis in understanding the relationships between different variables. By exploring the data and calculating descriptive statistics, I was able to gain insights into the impact of factors such as genre, duration, language, director, and budget on movie ratings.
The Five 'Whys approach helped me dig deeper into the problem and uncover the root causes behind certain relationships. By repeatedly asking "Why?", I was able to understand the underlying reasons why certain factors influence movie ratings.
Visualizations such as scatter plots and trendlines helped visualize the relationships between variables.
Overall, this project has taught me the importance of providing actionable insights that can drive decision-making for stakeholders. By analyzing the data and uncovering patterns and relationships, I can help movie producers, directors, and investors make informed decisions for their future projects.