IMDB Movie Analysis

Project 5

Ritika Chaudhary

Project Overview

The objective of this project is to analyze the IMDB movie dataset and identify factors that influence the success of movie, understand the relationship between budget and ratings. The analysis will focus on movie genre distribution, duration, language, directors with high IMDB scores and relation between profit and budget and how these all impact on IMDB scores

If we talk about the approaches used to analyze this project are:

- **Data Cleaning:** This step involves how we handled missing values, duplicates, converting data types.
- **Data Analysis:** Here we understand the relationship between different variables, correlation between movie ratings and other factors like genre, director, budget, etc.
- **Five 'Whys' Approach:** This technique helped to dig deeper into the problems.
    - Q: "Why do movies with higher budgets tend to have higher ratings?"
    - A: They can afford better production quality.
    - Q: "Why does better production quality lead to higher ratings?"
    - A: It enhances the viewer's experience.
    - Q: "Why does an enhanced viewer experience lead to higher ratings?"
    - A: Viewers are more likely to rate a movie highly if they enjoyed watching it.
    - Q: "Why are viewers more likely to rate a movie highly if they enjoyed watching it?"
    - A: Positive experiences lead to positive reviews.
    - Q: "Why do positive reviews matter?"
    - A: They influence other viewers' decisions to watch the movie, increasing its popularity and success.

Tech-Stack Used:

Microsoft Excel 2021

Purpose: data analysis, calculations and data visualisations.

Insights

Task 1: 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.
The following table represents the no. of movies for each genre.

genres	No. of Movies	average	Median	MAX	MIN	VAR	STDEV
Action	924	6.29	6.3	9	2.1	1.08	1.04
Adventure	752	6.45	6.6	8.9	2.3	1.26	1.12
Animation	194	6.70	6.8	8.6	2.8	0.98	0.99
Biography	237	7.16	7.2	8.9	4.5	0.48	0.69
Comedy	1443	6.19	6.3	8.8	1.9	1.07	1.03
Crime	692	6.55	6.6	9.3	2.4	0.97	0.99
Documentary	45	6.99	7.4	8.5	1.6	1.87	1.38
Drama	1846	6.80	6.9	9.3	2.1	0.79	0.89
Family	431	6.21	6.3	8.6	1.9	1.37	1.17
Fantasy	486	6.28	6.4	8.9	2.2	1.29	1.14
Film-Noir	1	7.70	7.7	7.7	7.7	0.00	0.00
History	146	7.15	7.2	8.9	5.5	0.44	0.67
Horror	374	5.90	5.9	8.6	2.3	0.98	0.99
Music	149	6.34	6.5	8.5	1.6	1.51	1.23
Musical	95	6.59	6.7	8.5	2.1	1.21	1.10
Mystery	371	6.48	6.5	8.6	3.1	1.01	1.00
Romance	839	6.43	6.5	8.5	2.1	0.92	0.96
Sci-Fi	479	6.33	6.4	8.8	1.9	1.36	1.17
Sport	143	6.59	6.8	8.3	2	1.09	1.05
Thriller	1074	6.37	6.4	9	2.7	0.94	0.97
War	149	7.06	7.1	8.6	4.3	0.64	0.80
Western	57	6.81	6.8	8.9	4.7	0.87	0.94

We separated the multiple genres for a single movie from a single column to multiple columns by using text to column function.

Formulas used for getting above insights:

=AVERAGE(IF($F$2:$M$3657=A3662, $U$2:$U$3657))
=MEDIAN(IF($F$2:$M$3657=A3662, $U$2:$U$3657))
=MAX(IF($F$2:$M$3657=A3662, $U$2:$U$3657))
=MIN(IF($F$2:$M$3657=A3662, $U$2:$U$3657))
=VAR.P(IF($F$2:$M$3657=A3662, $U$2:$U$3657))
=STDEV.S(IF($F$2:$M$3657=A3662, $U$2:$U$3657))

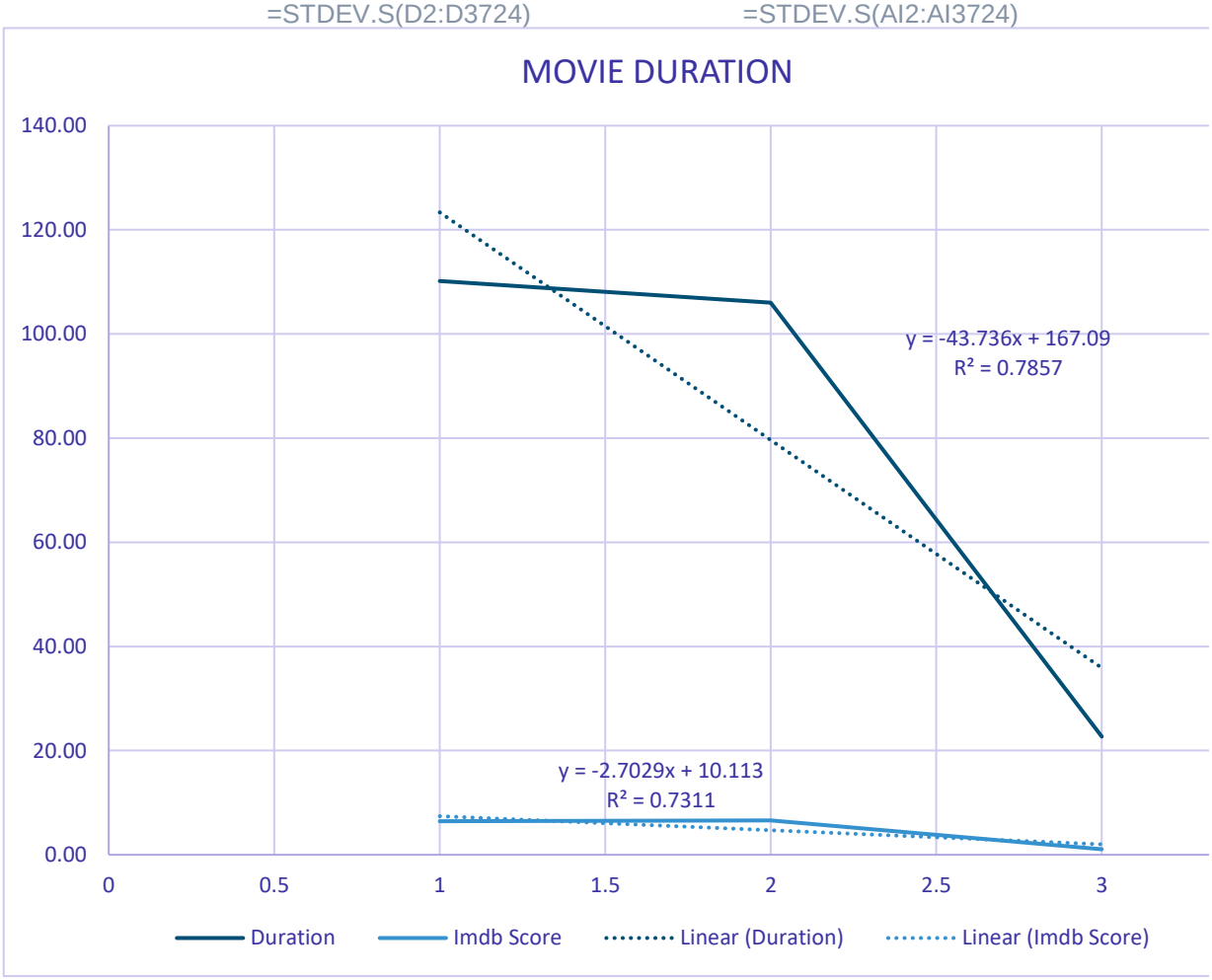
Task 2: Movie Duration Analysis

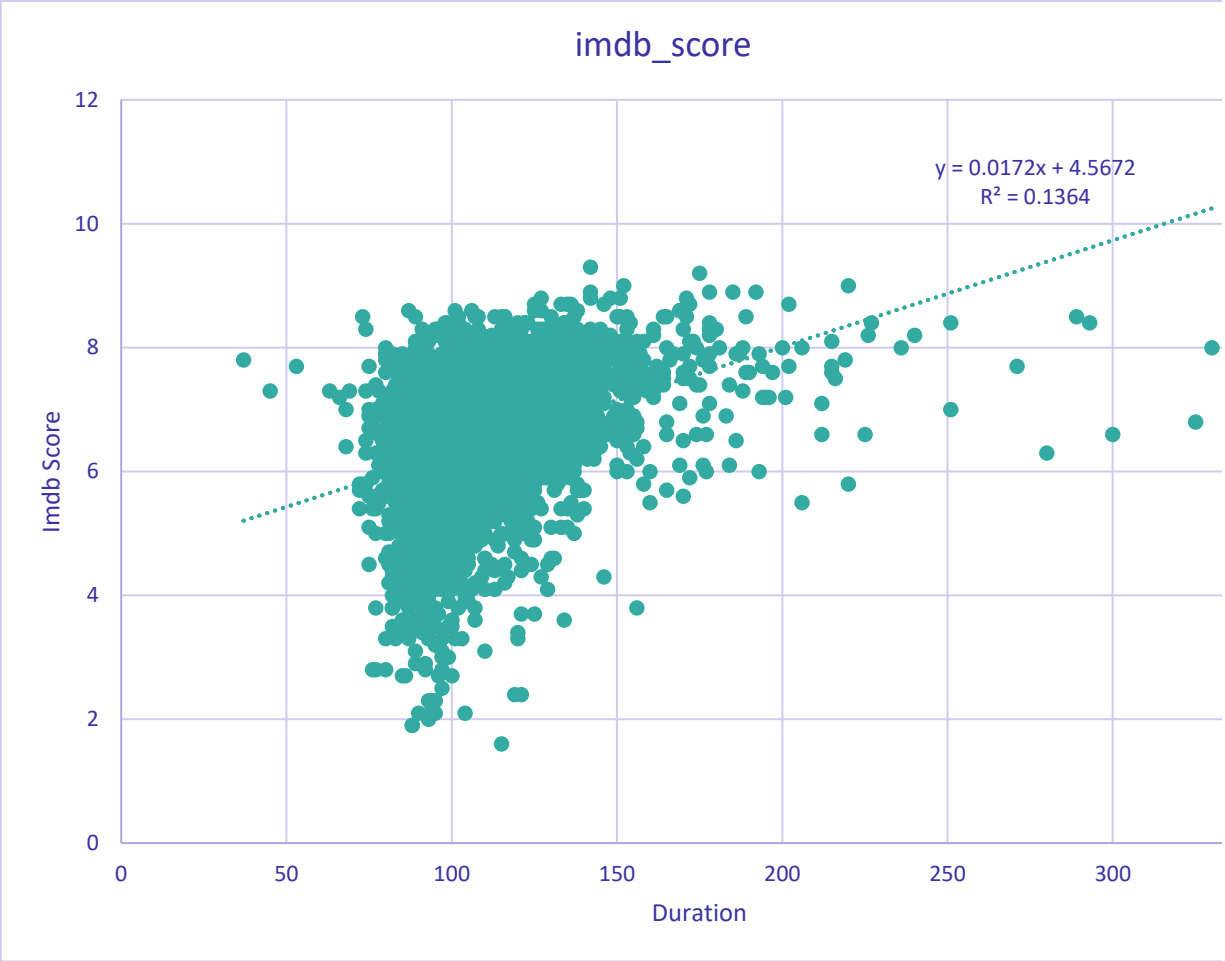
Analyse the distribution of movie durations and identify the relationship between movie duration and IMDB score.

Functions	Duration	Imdb Score
AVERAGE	110.26	6.47
MEDIAN	106.00	6.60
STD DEV	22.68	1.05

=AVERAGE(D2:D3724)
=MEDIAN(D2:D3724)
=STDEV.S(D2:D3724)

=AVERAGE(AI2:AI3724)
=MEDIAN(AI2:AI3724)
=STDEV.S(AI2:AI3724)





The slope is increasing as the movie duration increases IMDB scores tend to increase.

Task 3: Language Analysis

Examining the distribution of movies based on their language.

Created a list of unique languages in a separate column.
=UNIQUE(AB2:AB3724)

Then, statistics (Mean, Median and Standard Deviation)
=AVERAGEIFS(AI2:AI3725, AB2:AB3724, "Language")
=MEDIAN(IF(AB2:AB3724= "Language", AI2:AI3725))
=STDEV.S(IF(AB2:AB3724= "Language", AI2:AI3725))

language	Mean	Median	STD
English	6.43	6.50	1.05
Mandarin	7.02	7.25	0.77

French	7.36	7.30	0.52
Spanish	7.08	7.20	0.86
Cantonese	7.34	7.30	0.35
Maya	7.80	7.80	0.00
German	7.77	7.80	0.71
Hindi	7.22	7.40	0.80
Persian	8.13	8.40	0.55
Italian	7.19	7.00	1.16
Portuguese	7.76	8.00	0.98
Indonesian	7.90	7.90	0.42
Dari	7.50	7.50	0.14
Hebrew	8.00	8.00	0.00
Romanian	7.90	7.90	0.00
Danish	7.90	8.10	0.53
Aboriginal	6.95	6.95	0.78
Zulu	7.30	7.30	0.00
Arabic	7.20	7.20	0.00
Norwegian	7.15	7.30	0.57
Korean	7.70	7.70	0.57
Vietnamese	7.40	7.40	0.00
Japanese	7.66	8.00	0.99
Russian	6.50	6.50	0.00
Dutch	7.57	7.80	0.40
Bosnian	4.30	4.30	0.00
Mongolian	7.30	7.30	0.00
Kazakh	6.00	6.00	0.00
Aramaic	7.10	7.10	0.00
Filipino	6.70	6.70	0.00
Czech	7.40	7.40	0.00
Thai	6.63	6.60	0.45
Hungarian	7.10	7.10	0.00

Task 4: Director Analysis

Identifying top directors based on their imdb success.

=UNIQUE(B2:B3724)

=AVERAGEIFS($U$2:$U$3725, $B$2:$B$3724, "Director_Name")

Directors	Movie Ratings
Akira Kurosawa	8.70

Tony Kaye	8.60
Charles Chaplin	8.60
Ron Fricke	8.50
Majid Majidi	8.50
Damien Chazelle	8.50
Alfred Hitchcock	8.50
Sergio Leone	8.43
Christopher Nolan	8.43
Richard Marquand	8.40
Asghar Farhadi	8.40

Task 5: Budget Analysis

Exploring the relationship between movie budgets and their financial success.

Identified correlation coefficient between movie budgets and gross earnings =0.09831 which indicates a weak or no linear relationship.
Using the following function
=CORREL($R$2:$R$3724,$E$2:$E$3724)

Profit = Gross – Budget

Then, using max function for the profit column to find out movies with highest profit margins.
Top 10 movies with highest profit margins

Profit	Movie Titles
$523.506	AvatarÂ
$502.177	Jurassic WorldÂ
$458.672	TitanicÂ
$449.936	Star Wars: Episode IV - A New HopeÂ
$424.449	E.T. the Extra-TerrestrialÂ
$403.280	The AvengersÂ
$377.784	The Lion KingÂ
$359.545	Star Wars: Episode I - The Phantom MenaceÂ
$348.316	The Dark KnightÂ
$329.999	The Hunger GamesÂ

RESULT

This project fosters critical thinking, problem-solving and the ability to draw meaningful conclusions from data, all of which are crucial for a future career in data analytics. Additionally, the project's structured approach and use of tools like Excel enhance technical proficiency and provide a strong foundation for more advanced data analysis tasks. This project helped to develop skills in data preprocessing, visualization and statistical analysis.

DRIVE LINK
https://docs.google.com/spreadsheets/d/18H-jwOuunc_4UzYbG1_Wlc2U2PQJJdul/edit?usp=sharing&ouid=102610263604488941402&rtpof=true&sd=true

PRESENTATION LINK
https://drive.google.com/file/d/1T1tyjPWdPpf1w-ezK6ZxlS3EM1jaARXI/view?usp=sharing