IMDB MOVIE ANALYSIS

By: Nishtha Mehar

PROJECT DESCRIPTION

This project aims to analyze the factors influencing the success of movies on IMDB to provide actionable insights for stakeholders in the film industry. We will examine various attributes of the movies, such as genres, durations, languages, directors, and budgets, to understand their impact on IMDB ratings. The analysis will involve handling missing data, removing duplicates, and ensuring consistency in data formatting. We will calculate key statistics such as mean, median, and standard deviation to analyze genre distribution, explore the relationship between movie durations and ratings, and assess the influence of language on movie success. Additionally, we will identify top-performing directors based on their average IMDB scores and analyze the correlation between movie budgets and gross earnings while highlighting movies with the highest profit margins. The results will be presented with clear visualizations and actionable recommendations to enable better decision-making for movie producers, directors, and investors. This analysis will uncover patterns, identify trends, and offer insights to help optimize movie production strategies and improve the chances of creating successful films.

TECH-STACK USED

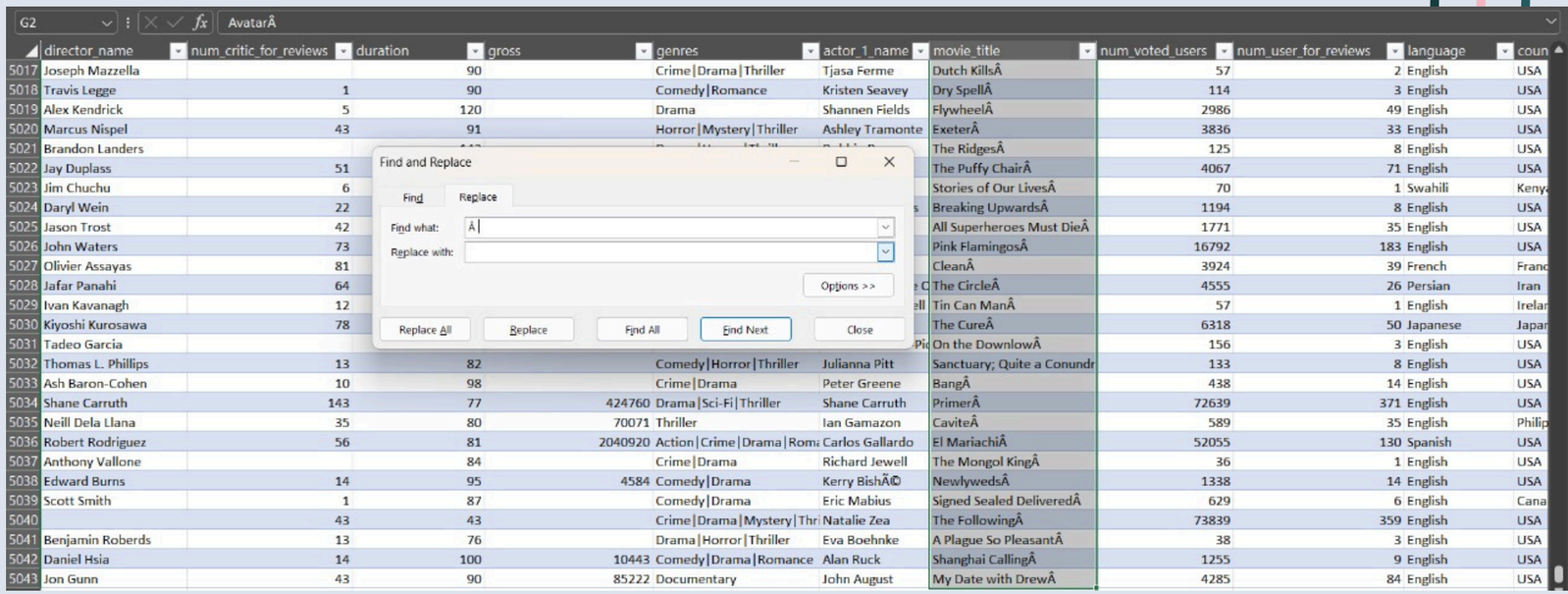
I used Microsoft Excel 365 for this analysis as it is a versatile and widely accessible tool for data manipulation, analysis, and visualization. Microsoft 365 offers a user-friendly interface with powerful features such as pivot tables, data filtering, and advanced functions for statistical analysis, which make it ideal for handling and summarizing large datasets. It also provides a wide range of charting options, allowing for clear and effective data visualization. Excel's ability to execute complex formulas and generate dynamic reports makes it an excellent choice for analyzing hiring data and extracting meaningful insights.

APPROACH

DATA WRANGLING

STEP 1: REMOVE IRRELEVANT COLUMNS

STEP 2: FILTER DATA WHICH HAS UNWANTED CHARACTERS



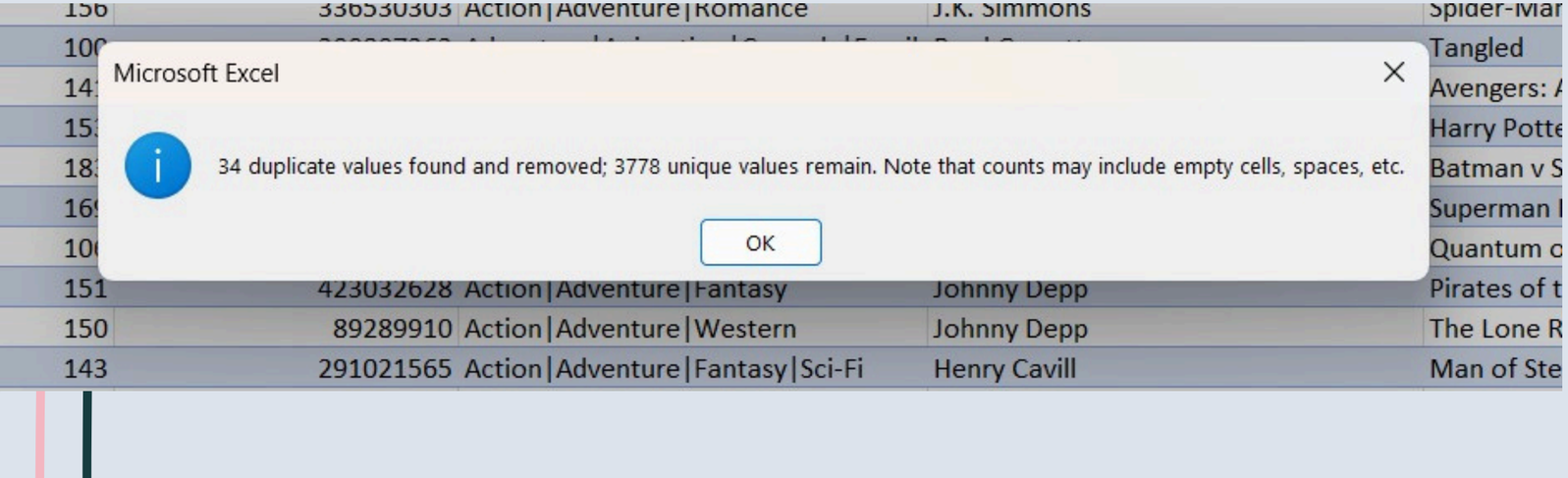
	director_name	num_critic_for_reviews	duration	gross	genres	actor_1_name	movie_title	num_voted_users	num_user_for_reviews	language	coun
5017	Joseph Mazzella		90		Crime|Drama|Thriller	Tjasa Ferme	Dutch KillsÂ	57	2	English	USA
5018	Travis Legge	1	90		Comedy|Romance	Kristen Seavey	Dry SpellÂ	114	3	English	USA
5019	Alex Kendrick	5	120		Drama	Shannen Fields	FlywheelÂ	2986	49	English	USA
5020	Marcus Nispel	43	91		Horror|Mystery|Thriller	Ashley Tramonte	ExeterÂ	3836	33	English	USA
5021	Brandon Landers						The RidgesÂ	125	8	English	USA
5022	Jay Duplass	51					The Puffy ChairÂ	4067	71	English	USA
5023	Jim Chuchu	6					Stories of Our LivesÂ	70	1	Swahili	Keny
5024	Daryl Wein	22					Breaking UpwardsÂ	1194	8	English	USA
5025	Jason Trost	42					All Superheroes Must DieÂ	1771	35	English	USA
5026	John Waters	73					Pink FlamingosÂ	16792	183	English	USA
5027	Olivier Assayas	81					CleanÂ	3924	39	French	Franc
5028	Jafar Panahi	64					The CircleÂ	4555	26	Persian	Iran
5029	Ivan Kavanagh	12					Tin Can ManÂ	57	1	English	Irelar
5030	Kiyoshi Kurosawa	78					The CureÂ	6318	50	Japanese	Japar
5031	Tadeo Garcia						On the DownlowÂ	156	3	English	USA
5032	Thomas L. Phillips	13	82		Comedy|Horror|Thriller	Julianna Pitt	Sanctuary; Quite a Conundr	133	8	English	USA
5033	Ash Baron-Cohen	10	98		Crime|Drama	Peter Greene	BangÂ	438	14	English	USA
5034	Shane Carruth	143	77	424760	Drama|Sci-Fi|Thriller	Shane Carruth	PrimerÂ	72639	371	English	USA
5035	Neill Dela Llana	35	80	70071	Thriller	Ian Gamazon	CaviteÂ	589	35	English	Philip
5036	Robert Rodriguez	56	81	2040920	Action|Crime|Drama|Roma	Carlos Gallardo	El MariachiÂ	52055	130	Spanish	USA
5037	Anthony Vallone		84		Crime|Drama	Richard Jewell	The Mongol KingÂ	36	1	English	USA
5038	Edward Burns	14	95	4584	Comedy|Drama	Kerry BishÃ©	NewlywedsÂ	1338	14	English	USA
5039	Scott Smith	1	87		Comedy|Drama	Eric Mabius	Signed Sealed DeliveredÂ	629	6	English	Cana
5040		43	43		Crime|Drama|Mystery|Thri	Natalie Zea	The FollowingÂ	73839	359	English	USA
5041	Benjamin Roberds	13	76		Drama|Horror|Thriller	Eva Boehnke	A Plague So PleasantÂ	38	3	English	USA
5042	Daniel Hsia	14	100	10443	Comedy|Drama|Romance	Alan Ruck	Shanghai CallingÂ	1255	9	English	USA
5043	Jon Gunn	43	90	85222	Documentary	John August	My Date with DrewÂ	4285	84	English	USA

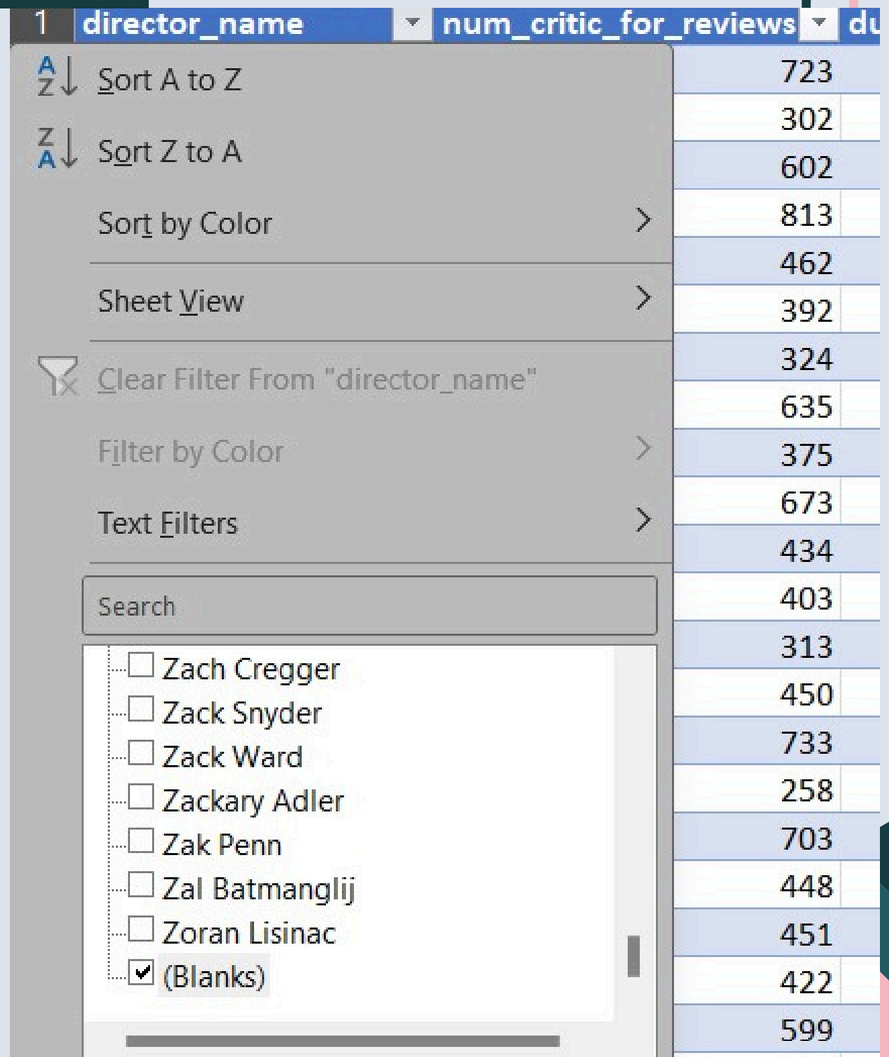
APPROACH

DATA WRANGLING

STEP 3: REMOVE ROWS WITH NULL VALUES

STEP 4: REMOVE DUPLICATES



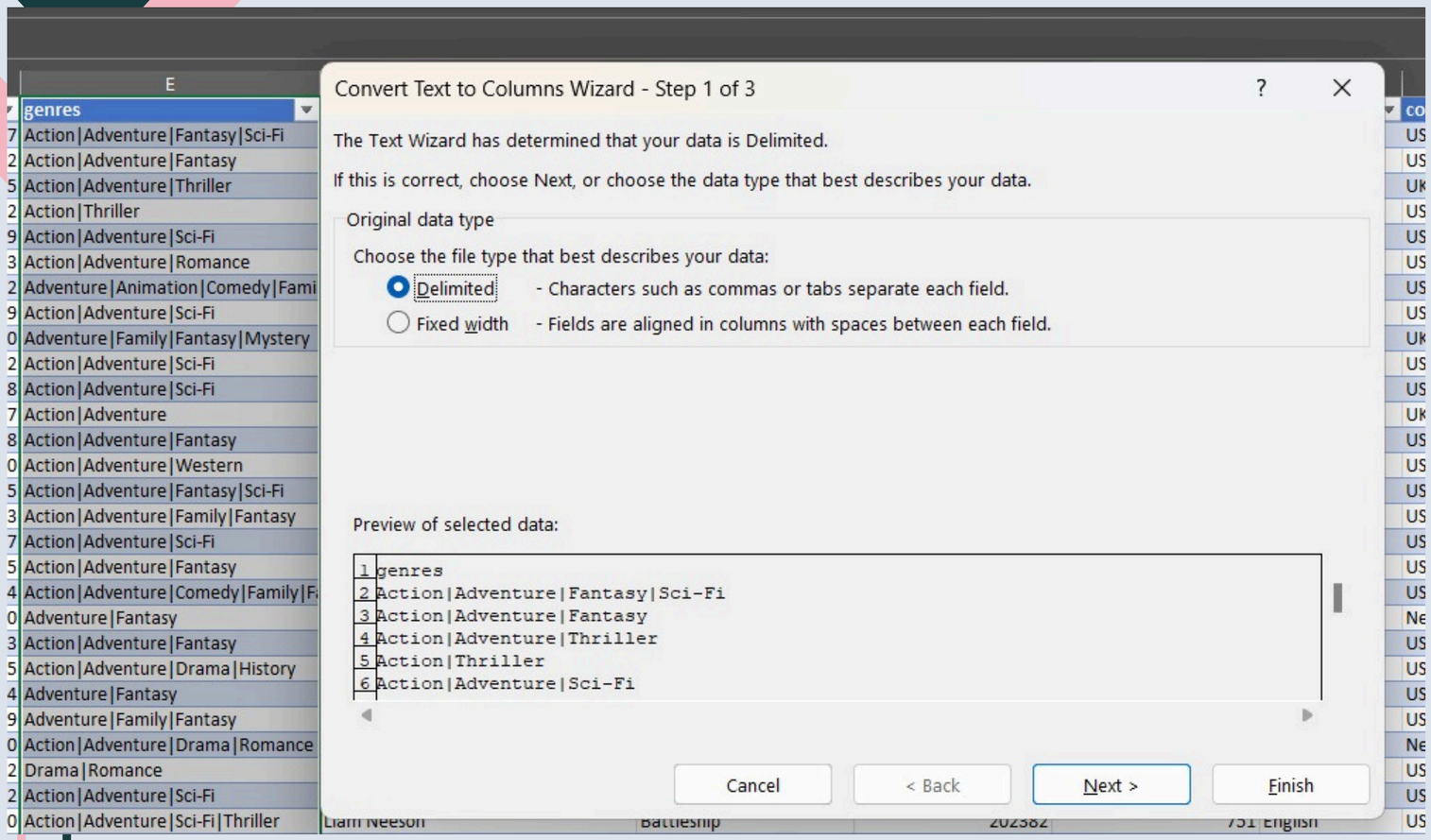


Movie Genre Analysis

APPROACH

1. Prepare the Dataset:
   - If the 'genres' column has multiple genres for a single movie (e.g., "Action|Adventure"), split them into separate rows using Excel's Text to Columns or Power Query.
2. Genre Distribution:
   - Use COUNTIF or use PIVOT TABLE to count the occurrences of each genre in the dataset .
3. Descriptive Statistics:
   - For each genre, filter the dataset by genre and calculate:
     - Mean (AVERAGE) of IMDB scores.
     - Median (MEDIAN) of IMDB scores.
     - Mode (MODE) of IMDB scores.
     - Range (MAX - MIN) of IMDB scores.
     - Variance (VAR.P or VAR.S).
     - Standard Deviation (STDEV.P or STDEV.S).
4. Comparison:
   - Compare the statistics across genres to observe trends.
   - Identify genres with consistently higher or lower IMDB scores.

INSIGHTS

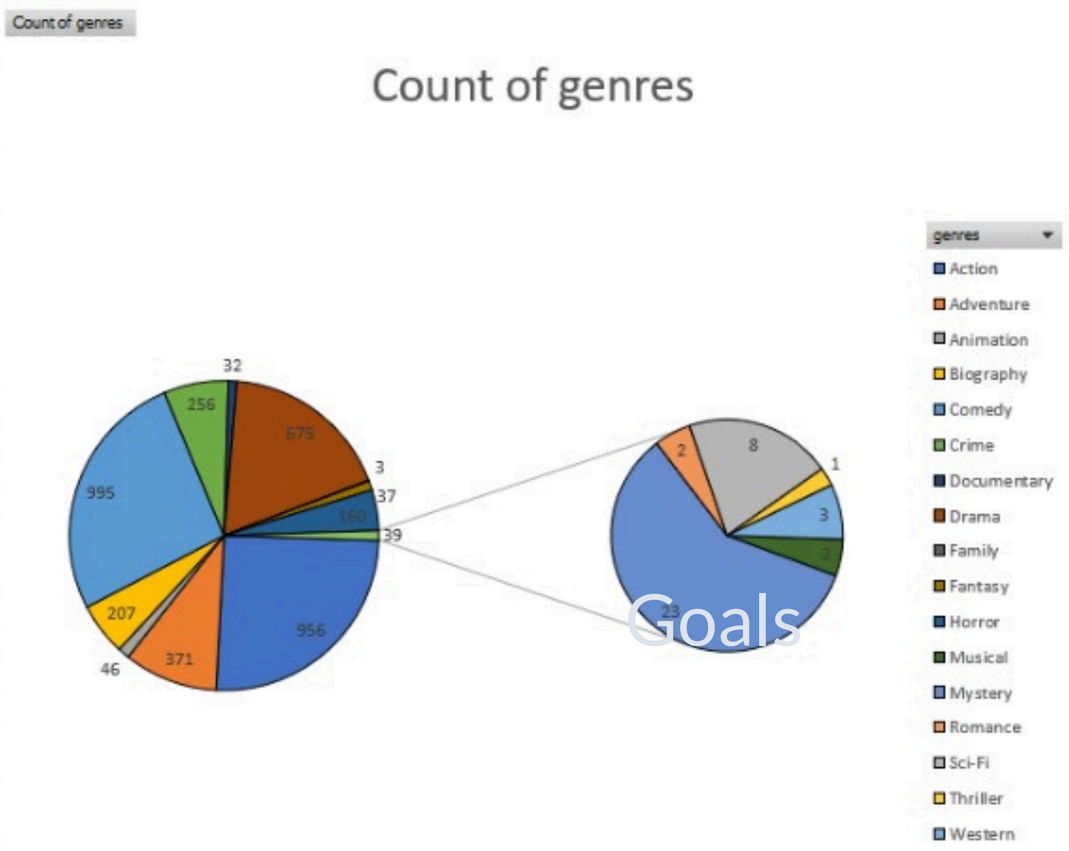


SELECT DELIMITER AND IT WILL SEPARATE INTO DIFFERENT COLUMNS AND THEN USE IT FOR GENRE COLUMN

INSIGHTS

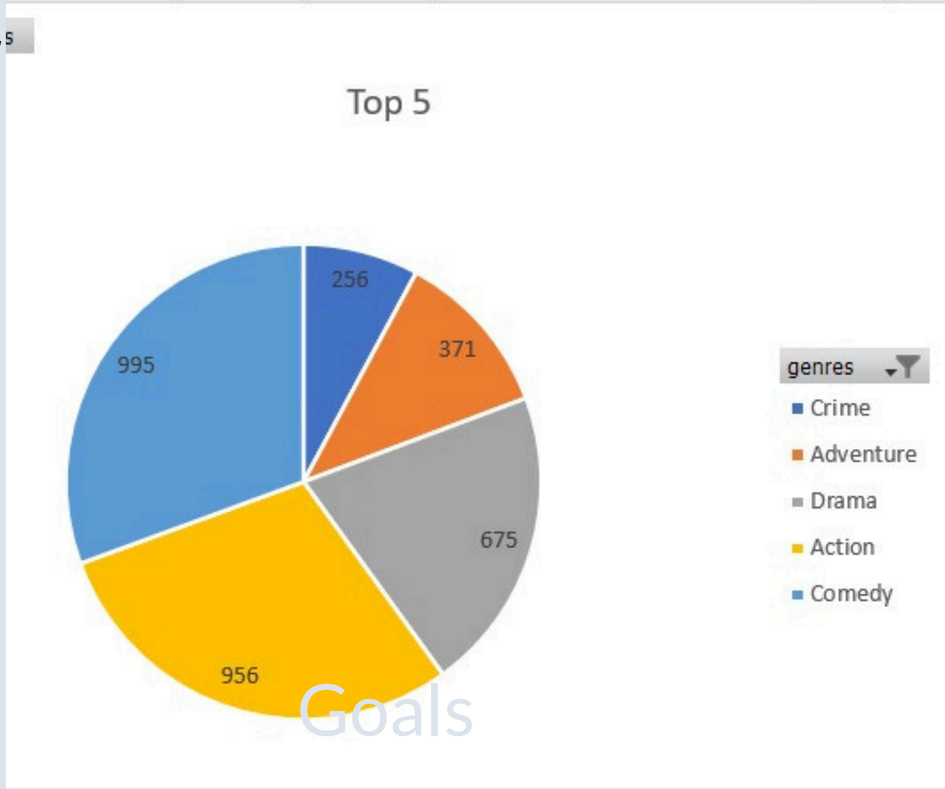
Row Labels	Count of genres
Action	956
Adventure	371
Animation	46
Biography	207
Comedy	995
Crime	256
Documentary	32
Drama	675
Family	3
Fantasy	37
Horror	160
Musical	2
Mystery	23
Romance	2
Sci-Fi	8
Thriller	1
Western	3
Grand Total	**3777**

Genre Manipulation	
Median	37
Max	995
Min	1
Mode	3
Var	112323.7794
Std	335.1473995



INSIGHTS

1. Mode (3):
   - This highlights the presence of several underrepresented genres with only a small number of movies (e.g., Family, Musical, Romance, Thriller). These genres may not be primary choices for filmmakers, potentially due to niche audiences or lower market demand.
2. Median (37):
   - The median is the middle value when the data is ordered. Here, the median genre count is 37, meaning that half of the genres have fewer than 37 movies, and half have more.
   - Insight: Most genres are on the lower end of the distribution, as the median is much lower than the highest genre count (995). This indicates that a few genres dominate production while many are underrepresented.
3. Maximum (995):
   - The maximum represents the genre with the highest count of movies, which is Comedy.
   - Insight: Comedy's dominance (995 movies) suggests it is a highly preferred genre, likely due to its broad appeal and commercial viability.
4. Minimum (1):
   - The minimum is the lowest genre count, which is 1 (e.g., Thriller).
   - Insight: This reflects genres that are either experimental, region-specific, or less popular among both filmmakers and audiences.
5. Variance (112,323.78):
   - Variance measures the spread of the data points from the mean. A high variance indicates significant differences in genre counts.
   - Insight: The large variance shows that genre counts are highly dispersed, with some genres (like Comedy and Action) far outpacing others (like Thriller or Romance).
6. Standard Deviation (335.15):
   - Standard deviation represents the average distance of each data point from the mean. A high standard deviation here (335.15) supports the idea of uneven genre distribution.
   - Insight: The genre counts are not evenly distributed. Only a few genres are highly popular, while others are scarcely represented.



Movie Duration Analysis

APPROACH

- Descriptive Statistics for Duration:
- Calculate:
  - Mean (AVERAGE) of movie durations.
  - Median (MEDIAN) of movie durations.
  - Standard Deviation (STDEV.P or STDEV.S).
- Scatter Plot:

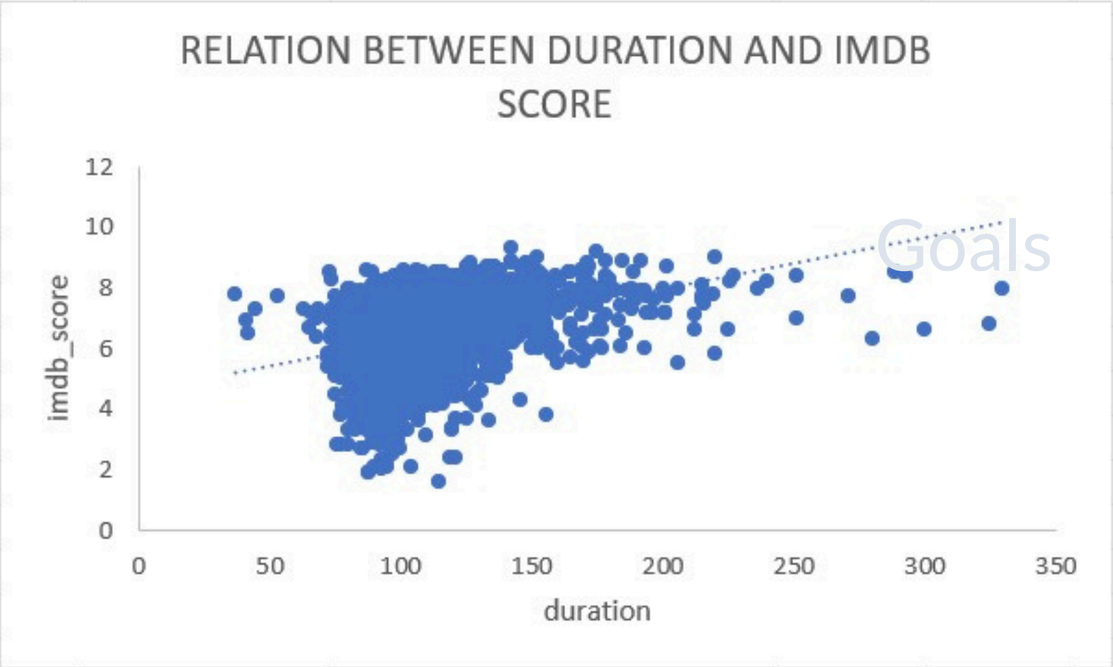
  - Create a scatter plot with movie duration on the X-axis and IMDB score on the Y-axis.
- Add a trendline using Excel's Add Trendline feature.
- Analyze the trendline's slope to understand the relationship.

INSIGHTS

duration	imdb_score
178	7.9
169	7.1
148	6.8
164	8.5
132	6.6
156	6.2
100	7.8
141	7.5
153	7.5
183	6.9
169	6.1
106	6.7
151	7.3
150	6.5
143	7.2
150	6.6
173	8.1
136	6.7
106	6.8
164	7.5
153	7
156	6.7
186	7.9
113	6.1
201	7.2
194	7.7

DURATION STATISTICS	
Mean	110.1837437
Median	106
Std	22.77240051

WHENEVER WE COMPARE 2 VARIABLES WE USE SCATTER CHART



INSIGHTS

SUMMARY STATISTICS:

Mean Duration (110.18 minutes):

- The average movie duration is approximately 110 minutes, aligning with the industry standard for feature films. This suggests most movies in the dataset fall into the conventional runtime range.

Median Duration (106 minutes):

- Half of the movies are shorter than 106 minutes, indicating a slight skew towards shorter films.

Standard Deviation (22.77 minutes):

- The duration data is moderately spread, meaning most movies have runtimes within approximately ±23 minutes of the mean (i.e., between 87 and 133 minutes).

SCATTER PLOT INSIGHTS:

- The scatter plot shows a weak positive correlation between movie duration and IMDB scores. This implies:
  - Longer movies tend to have slightly higher IMDB scores, but the trend is not strong or consistent.
  - Most movies cluster around durations of 80–130 minutes with scores between 6 and 8, showing that a majority of movies fall within the average runtime and rating.

POTENTIAL FACTORS INFLUENCING THE TREND:

Audience Engagement:

- Movies with optimal runtimes (90–130 minutes) may balance storytelling with viewer attention spans, leading to higher ratings. Very short or excessively long movies might fail to maintain audience interest, lowering scores.

Genre Influence:

- Certain genres (e.g., action, drama) might contribute to longer durations and higher scores, while shorter runtimes might align with less complex genres like comedy or animation.

Language Analysis

3	Row Labels	Count of language
4	English	3604
5	French	36
6	Spanish	23
7	Mandarin	14
8	German	13
9	Japanese	12
10	Cantonese	8
11	Hindi	8
12	Italian	7
13	Portuguese	5
14	Korean	5
15	Norwegian	4
16	Dutch	3
17	Thai	3
18	Persian	3
19	Danish	3
20	Hebrew	2
21	Aboriginal	2
22	Indonesian	2
23	Dari	2
24	Bosnian	1
25	Russian	1
26	Arabic	1
27	Kazakh	1
28	Swedish	1
29	Aramaic	1
30	Zulu	1
31	Dzongkha	1
32	Romanian	1
33	Maya	1
34	Filipino	1
35	Mongolian	1
36	Telugu	1
37	None	1
38	Vietnamese	1
39	Czech	1
40	Hungarian	1
41	Icelandic	1
42	(blank)	
43	**Grand Total**	**3777**

APPROACH

Language Distribution:

- Use COUNTIF to count the number of movies for each language.

Descriptive Statistics:

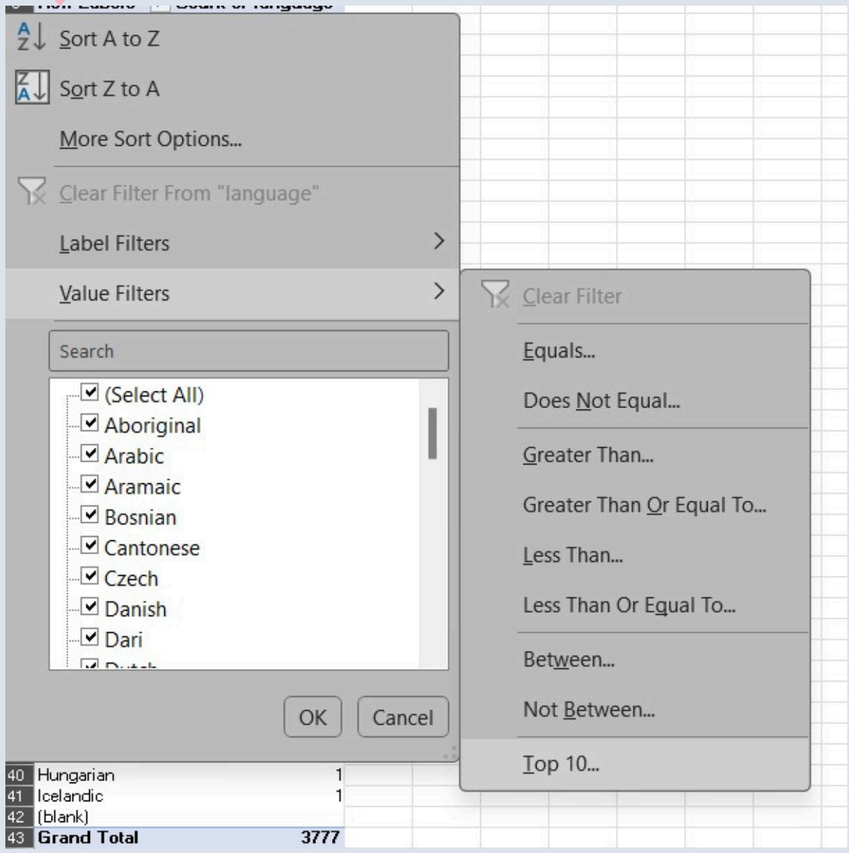
- For each language, filter the dataset and calculate:
- Mean (AVERAGE) of IMDB scores.
- Median (MEDIAN) of IMDB scores.
- Standard Deviation (STDEV.P or STDEV.S).

Comparison:

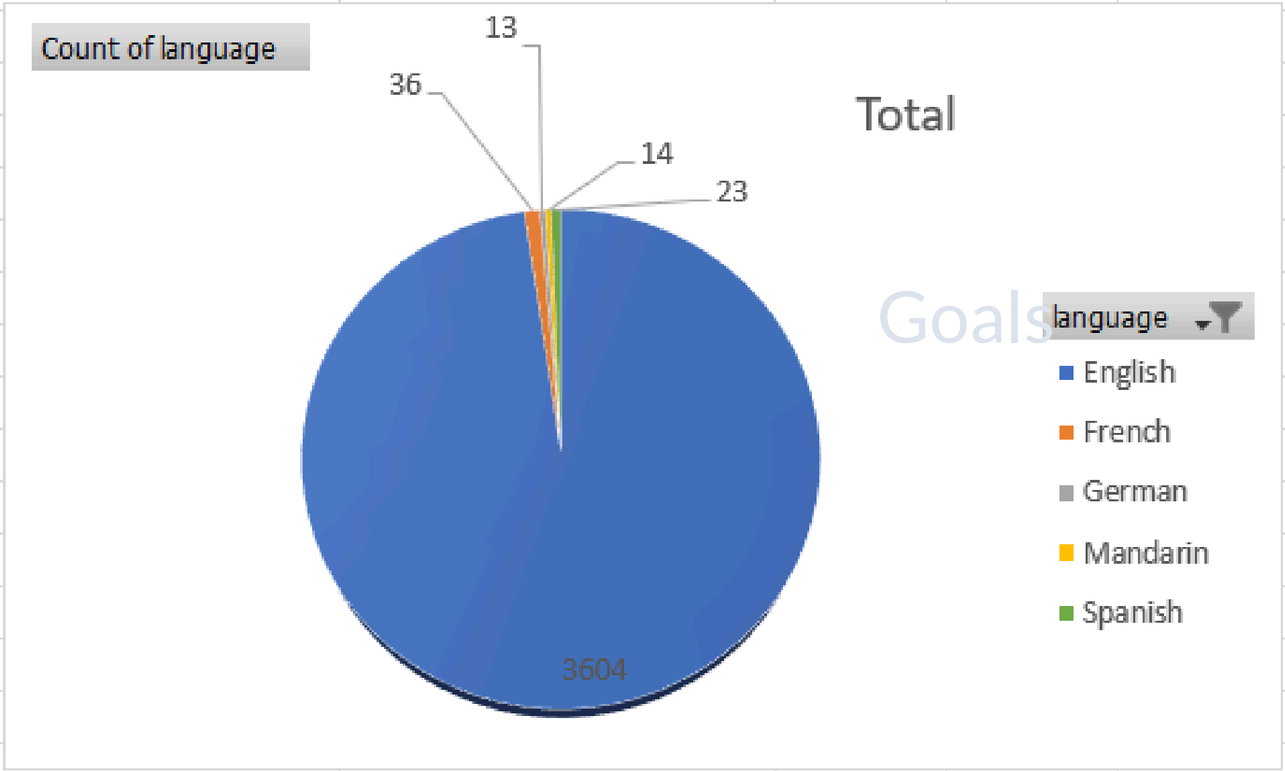
- Compare statistics to identify languages associated with higher or lower ratings.
- Identify if specific languages correlate with higher IMDB scores.

INSIGHTS

TOP 5 COMMON LANGUAGE



Row Labels	Count of language
English	3604
French	36
German	13
Mandarin	14
Spanish	23
Grand Total	**3690**



INSIGHTS

H	I	J	K	L
language	MEAN	MEDIAN	STD	
English	6.43	6.5	1.05	
Mandarin	7.02	7.25	0.77	
Aboriginal	6.95	6.95	0.78	
Spanish	7.08	7.2	0.86	
French	7.30	7.25	0.57	
Filipino	6.70	6.7	0.00	
Maya	7.80	7.8	0.00	
Kazakh	6.00	6	0.00	
Telugu	8.40	8.4	0.00	
Cantonese	7.24	7.3	0.44	
Japanese	7.63	7.8	0.90	
Aramaic	7.10	7.1	0.00	
Italian	7.19	7	1.16	
Dutch	7.57	7.8	0.40	
Dari	7.50	7.5	0.14	
German	7.69	7.7	0.64	
Mongolian	7.30	7.3	0.00	
Thai	6.63	6.6	0.45	
Bosnian	4.30	4.3	0.00	
Korean	7.70	7.7	0.57	
Hungarian	7.10	7.1	0.00	
Hindi	7.18	7.3	0.76	
Icelandic	6.90	6.9	0.00	
Danish	7.90	8.1	0.53	
Portuguese	7.76	8	0.98	
Norwegian	7.15	7.3	0.57	

2		
3	Row Labels	Count of movie_title
4	Aboriginal	2
5	Arabic	1
6	Aramaic	1
7	Bosnian	1
8	Cantonese	8
9	Czech	1
10	Danish	3
11	Dari	2
12	Dutch	3
13	Dzongkha	1
14	English	3604
15	Filipino	1
16	French	36
17	German	13
18	Hebrew	2
19	Hindi	8
20	Hungarian	1
21	Icelandic	1
22	Indonesian	2
23	Italian	7
24	Japanese	12
25	Kazakh	1
26	Korean	5
27	Mandarin	14
28	Maya	1
29	Mongolian	1
30	None	1
31	Norwegian	4
32	Persian	3
33	Portuguese	5
34	Romanian	1
35	Russian	1
36	Spanish	23
37	Swedish	1
38	Telugu	1
39	Thai	3
40	Vietnamese	1
41	Zulu	1
42	Grand Total	3777
43		

Mean (Average):

- **Languages like Telugu (8.40), Danish (7.90),Persian(8.13) have the highest average IMDB scores, suggesting movies in these languages tend to perform exceptionally well.**
- **Bosnian (4.30) has the lowest average score, indicating poor performance for movies in this language.**

Median:

- **The median reflects the "middle score" and confirms trends seen in the mean. For example, Telugu (8.40) and Danish (8.1) have high medians, reinforcing their strong performance.**
- **Bosnian (4.30) remains an outlier with low ratings across the board.**

Standard Deviation (STD):

- **A low standard deviation (e.g., Mongolian (0.00) and Maya (0.00)) indicates consistent IMDB scores across movies in these languages, suggesting reliable quality.**
- **A high standard deviation (e.g., Italian (1.16) and English (1.05)) shows more variation in movie ratings, which could indicate a mix of both highly successful and poorly received movies.**

Director Analysis

APPROACH

Average IMDB Score by Director:

- Group the dataset by directors.
- Calculate the average (AVERAGE) IMDB score for each director using =AVERAGEIF(A:A,A4,B:B)

Percentile Calculation:

- Use PERCENTILE.EXC to find percentiles of the directors' average scores like =PERCENTILE.EXC(B:B,0.9)

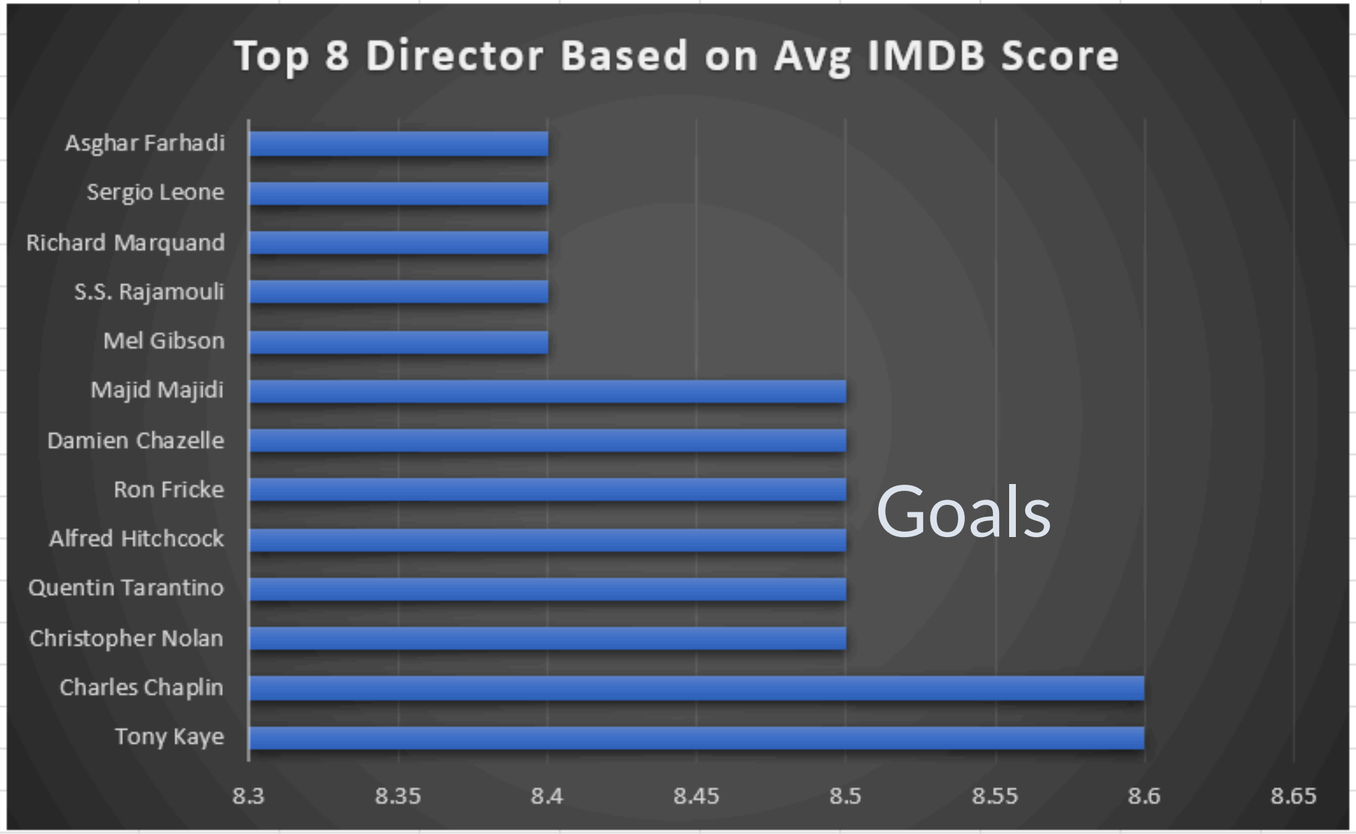
Comparison:

- Identify directors in the top percentile and compare their scores with the dataset's overall score distribution.

INSIGHTS

director_name	imdb_score	avg
Tony Kaye	8.6	8.6
Charles Chaplin	8.6	8.6
Christopher Nolan	8.5	8.5
Quentin Tarantino	8.5	8.5
Alfred Hitchcock	8.5	8.5
Ron Fricke	8.5	8.5
Damien Chazelle	8.5	8.5
Majid Majidi	8.5	8.5
Mel Gibson	8.4	8.4
S.S. Rajamouli	8.4	8.4
Richard Marquand	8.4	8.4
Sergio Leone	8.4	8.4
Asghar Farhadi	8.4	8.4

90 percentile
7.5



INSIGHTS

Top Directors Based on IMDB Scores:

- The chart highlights Tony Kaye and Charles Chaplin as the top directors, both achieving an average IMDB score of 8.6, closely followed by Christopher Nolan, Quentin Tarantino, and Alfred Hitchcock with a score of 8.5.
- These high scores suggest these directors consistently produce critically acclaimed movies that resonate with audiences.

Percentile Calculation:

- The 90th percentile IMDB score is 7.5, indicating that the top-performing directors far exceed this threshold, solidifying their positions as leaders in the industry.
- Directors who fall above this percentile are considered exceptional in terms of delivering quality films.

Comparison to Distribution:

- The average IMDB score for these directors (8.4–8.6) is significantly higher than the general distribution of scores in the dataset (where the mean is often closer to 6–7).
- This shows that the selected directors are in an elite group, as their movies consistently achieve higher ratings.

Budget Analysis

APPROACH

CORRELATION COEFFICIENT:

- Use CORREL to calculate the correlation coefficient between the movie budget and gross earnings.(=CORREL(C:C,B:B))

PROFIT MARGIN:

- Add a new column for Profit Margin: Gross Earnings - Budget.(=B3-C3)
- Use MAX to identify the movies with the highest profit margins. (=MAX(D:D))

INSIGHTS

	B	C	D	E	F
1	gross	budget	Profit Margin		highest profit margin
2	760505847	237000000	523505847		523505847
3	652177271	150000000	502177271		
4	658672302	200000000	458672302		
5	460935665	11000000	449935665		correlation function
6	434949459	10500000	424449459		0.095158696
7	623279547	220000000	403279547		
8	422783777	45000000	377783777		
9	474544677	115000000	359544677		
10	533316061	185000000	348316061		
11	407999255	78000000	329999255		
12	363024263	58000000	305024263		
13	424645577	130000000	294645577		
14	356784000	63000000	293784000		
15	368049635	76000000	292049635		
16	350123553	58800000	291323553		
17	380838870	94000000	286838870		
18	436471036	150000000	286471036		
19	377019252	94000000	283019252		
20	309125409	32500000	276625409		
21	329691196	55000000	274691196		
22	290158751	18000000	272158751		
23	285761243	18000000	267761243		
24	380262555	113000000	267262555		
25	403706375	139000000	264706375		
26	336029560	74000000	262029560		

CORRELATION BETWEEN BUDGET AND GROSS EARNINGS:

- Correlation Value: 0.0951 (calculated using Excel's CORREL function).
- Interpretation: This value indicates a very weak positive correlation between movie budgets and gross earnings. It suggests that higher budgets do not necessarily guarantee higher gross earnings. Other factors, such as storyline, star cast, marketing, or release timing, might play a more significant role in financial success.

PROFIT MARGIN ANALYSIS:

- Profit Margin Formula: Profit Margin = Gross Earnings - Budget.
- Highest Profit Margin: 523,505,847 (calculated using Excel's MAX function).
- This shows the movie with the highest financial gain relative to its budget. High profit margins often signify effective cost management and strong audience appeal.

LINK

EXCEL LINK

https://docs.google.com/spreadsheets/d/14K-qz-M7i_XbUOUabGVGNCu6fLti31sh/edit?usp=sharing&ouid=111854720256815218942&rtpof=true&sd=true

LOOM VIDEO PRESENTATION

LINhttps://www.loom.com/share/db7c645b7c764806b6af06d325fa43d7?sid=a9269c6d-7b92-4643-8128-696fde865e56K:

CONCLUSION

Working on this analysis allowed me to dive deep into various factors influencing the success of a movie on IMDB. Each task provided valuable insights into how genres, duration, language, directors, and budgets affect movie ratings. The project helped me understand the complexities of the movie industry and the importance of leveraging data for decision-making.

The ability to analyze and visualize data in Excel gave me hands-on experience with real-world data, improving my skills in data manipulation, statistical analysis, and data storytelling. By using descriptive statistics, correlation analysis, and visualization techniques, I was able to uncover patterns and trends that provide actionable insights for movie producers, directors, and investors.

This analysis not only shed light on what factors contribute to a movie's success but also demonstrated the power of data-driven decision-making. The insights gained from this project can assist stakeholders in making informed choices about future movie productions, improving their chances of success in a highly competitive industry.

THANK YOU