IMDB Movie Analysis

Project 5

Project Description

- The goal of this project is to investigate the factors that influence the success of movies on IMDB. Success is defined by high IMDB ratings, which reflect audience and critical reception. This analysis is significant for movie producers, directors, and investors who seek to understand the elements contributing to a movie's success to make informed decisions for future projects.
- My goal with this project is to provide actionable insights into the key factors that influence a movie's success on IMDB. These findings will benefit the film industry by offering data-driven strategies for producing and marketing movies that resonate with audiences and critics.

Approach

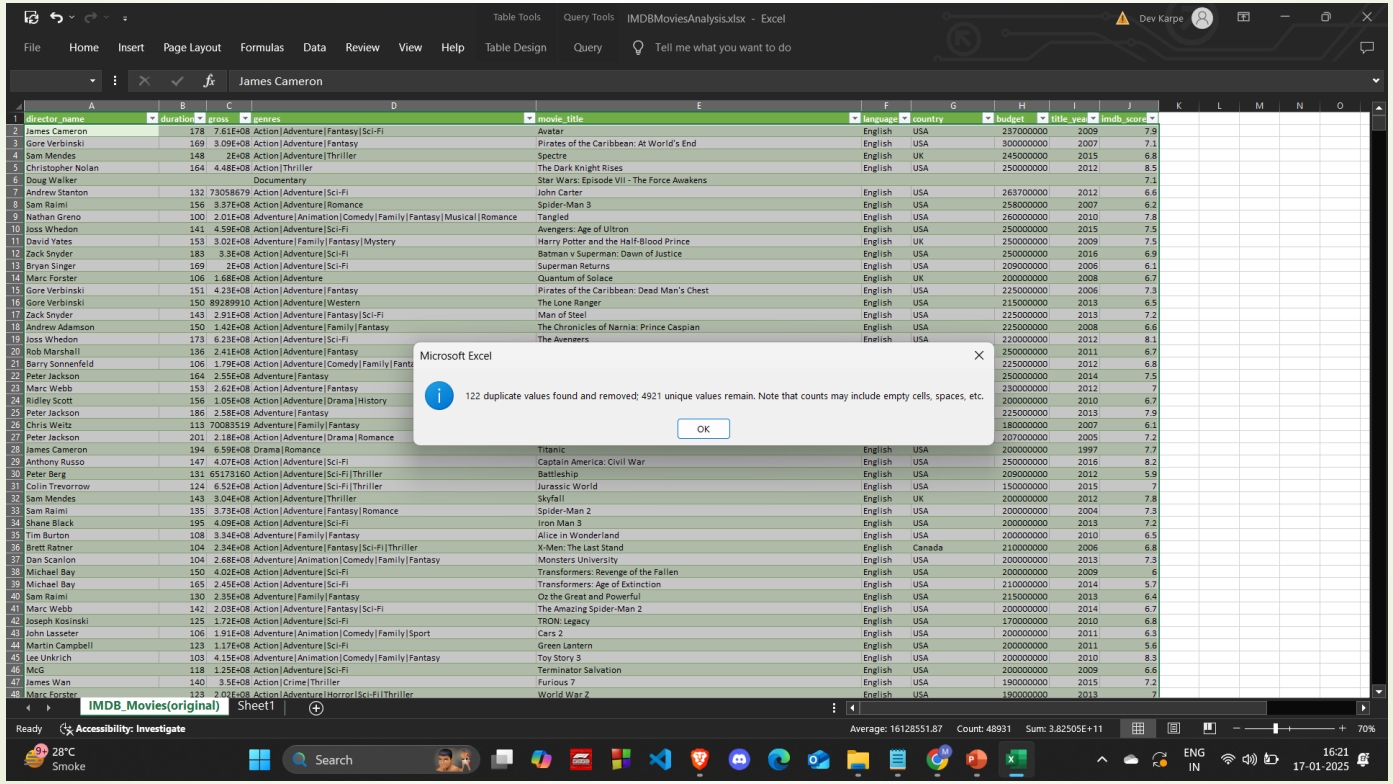
- Firstly, I downloaded the dataset from the given source.
- I then converted the .csv file to a .xlsx file
- Next I gained an understanding of the various columns of the dataset.
- This was followed by the process of cleaning the data to get rid of inconsistencies.
- Data cleaning involved handling missing data, removing duplicate columns, correcting spelling mistakes, filling in some empty cells and deleting unnecessary columns that would not be needed for the analysis.

Tech Stack Used

- Microsoft Excel 2019 was used to perform the analysis
- Microsoft PowerPoint 2019 was used to display the project findings

Insights

- **Data cleaning:**
- Firstly, I removed duplicate values



director_name	duration	gross	genres	movie_title	language	country	budget	title_year	imdb_score
James Cameron	178	7.61E+08	Action|Adventure|Fantasy|Sci-Fi	Avatar	English	USA	237000000	2009	7.9
Gore Verbinski	169	3.09E+08	Action|Adventure|Fantasy	Pirates of the Caribbean: At World's End	English	USA	300000000	2007	7.1
Sam Mendes	148	2E+08	Action|Adventure|Thriller	Spectre	English	UK	245000000	2015	6.8
Christopher Nolan	164	4.48E+08	Action|Thriller	The Dark Knight Rises	English	USA	250000000	2012	8.5
Doug Walker			Documentary	Star Wars: Episode VII - The Force Awakens					7.1
Andrew Stanton	132	73058679	Action|Adventure|Sci-Fi	John Carter	English	USA	263700000	2012	6.6
Sam Raimi	156	3.37E+08	Action|Adventure|Romance	Spider-Man 3	English	USA	258000000	2007	6.2
Nathan Greno	100	2.01E+08	Adventure|Animation|Comedy|Family|Fantasy|Musical|Romance	Tangled	English	USA	260000000	2010	7.8
Joss Whedon	141	4.59E+08	Action|Adventure|Sci-Fi	Avengers: Age of Ultron	English	USA	250000000	2015	7.5
David Yates	153	3.02E+08	Adventure|Family|Fantasy|Mystery	Harry Potter and the Half-Blood Prince	English	UK	250000000	2009	7.5
Zack Snyder	183	3.3E+08	Action|Adventure|Sci-Fi	Batman v Superman: Dawn of Justice	English	USA	250000000	2016	6.9
Bryan Singer	169	2E+08	Action|Adventure|Sci-Fi	Superman Returns	English	USA	209000000	2006	6.1
Marc Forster	106	1.68E+08	Action|Adventure	Quantum of Solace	English	UK	200000000	2008	6.7
Gore Verbinski	151	4.23E+08	Action|Adventure|Fantasy	Pirates of the Caribbean: Dead Man's Chest	English	USA	225000000	2006	7.3
Gore Verbinski	150	89289910	Action|Adventure|Western	The Lone Ranger	English	USA	215000000	2013	6.5
Zack Snyder	143	2.91E+08	Action|Adventure|Fantasy|Sci-Fi	Man of Steel	English	USA	225000000	2013	7.2
Andrew Adamson	150	1.42E+08	Action|Adventure|Family|Fantasy	The Chronicles of Narnia: Prince Caspian	English	USA	225000000	2008	6.6
Joss Whedon	173	6.23E+08	Action|Adventure|Sci-Fi	The Avengers	English	USA	220000000	2012	8.1
Rob Marshall	136	2.41E+08	Action|Adventure|Fantasy				250000000	2011	6.7
Barry Sonnenfeld	106	1.79E+08	Action|Adventure|Comedy|Family|Fanta				225000000	2012	6.8
Peter Jackson	164	2.55E+08	Adventure|Fantasy				250000000	2014	7.5
Marc Webb	153	2.62E+08	Action|Adventure|Fantasy				230000000	2012	7
Ridley Scott	156	1.05E+08	Action|Adventure|Drama|History				200000000	2010	6.7
Peter Jackson	186	2.58E+08	Adventure|Fantasy				225000000	2013	7.9
Chris Weitz	113	70083519	Adventure|Family|Fantasy				180000000	2007	6.1
Peter Jackson	201	2.18E+08	Action|Adventure|Drama|Romance				207000000	2005	7.2
James Cameron	194	6.59E+08	Drama|Romance	Titanic	English	USA	200000000	1997	7.7
Anthony Russo	147	4.07E+08	Action|Adventure|Sci-Fi	Captain America: Civil War	English	USA	250000000	2016	8.2
Peter Berg	131	65173160	Action|Adventure|Sci-Fi|Thriller	Battleship	English	USA	209000000	2012	5.9
Colin Trevorrow	124	6.52E+08	Action|Adventure|Sci-Fi|Thriller	Jurassic World	English	USA	150000000	2015	7
Sam Mendes	143	3.04E+08	Action|Adventure|Thriller	Skyfall	English	UK	200000000	2012	7.8
Sam Raimi	135	3.73E+08	Action|Adventure|Fantasy|Romance	Spider-Man 2	English	USA	200000000	2004	7.3
Shane Black	195	4.09E+08	Action|Adventure|Sci-Fi	Iron Man 3	English	USA	200000000	2013	7.2
Tim Burton	108	3.34E+08	Adventure|Family|Fantasy	Alice in Wonderland	English	USA	200000000	2010	6.5
Brett Ratner	104	2.34E+08	Action|Adventure|Fantasy|Sci-Fi|Thriller	X-Men: The Last Stand	English	Canada	210000000	2006	6.8
Dan Scanlon	104	2.68E+08	Adventure|Animation|Comedy|Family|Fantasy	Monsters University	English	USA	200000000	2013	7.3
Michael Bay	150	4.02E+08	Action|Adventure|Sci-Fi	Transformers: Revenge of the Fallen	English	USA	200000000	2009	6
Michael Bay	165	2.45E+08	Action|Adventure|Sci-Fi	Transformers: Age of Extinction	English	USA	210000000	2014	5.7
Sam Raimi	130	2.35E+08	Adventure|Family|Fantasy	Oz the Great and Powerful	English	USA	215000000	2013	6.4
Marc Webb	142	2.03E+08	Action|Adventure|Fantasy|Sci-Fi	The Amazing Spider-Man 2	English	USA	200000000	2014	6.7
Joseph Kosinski	125	1.72E+08	Action|Adventure|Sci-Fi	TRON: Legacy	English	USA	170000000	2010	6.8
John Lasseter	106	1.91E+08	Adventure|Animation|Comedy|Family|Sport	Cars 2	English	USA	200000000	2011	6.3
Martin Campbell	123	1.17E+08	Action|Adventure|Sci-Fi	Green Lantern	English	USA	200000000	2011	5.6
Lee Unkrich	103	4.15E+08	Adventure|Animation|Comedy|Family|Fantasy	Toy Story 3	English	USA	200000000	2010	8.3
McG	118	1.25E+08	Action|Adventure|Sci-Fi	Terminator Salvation	English	USA	200000000	2009	6.6
James Wan	140	3.5E+08	Action|Crime|Thriller	Furious 7	English	USA	190000000	2015	7.2
Marc Forster	123	2.02E+08	Action|Adventure|Horror|Sci-Fi|Thriller	World War Z	English	USA	190000000	2013	7

- Next, the rows containing empty cells were deleted
- I then changed the movie_title column to correct the spelling mistakes it had. This was done by creating a new column and using the substitute function.
- Unnecessary columns were deleted. The columns color, director_facebook_likes, actor_3_facebook_likes, actor_2_name, actor_1_facebook_likes, actor_1_name, cast_total_facebook_likes, actor_3_name, facenumber_in_poster, plot_keywords, actor_2_facebook_likes, aspect_ratio, movie_imdb_link, content_rating were irrelevant and that's why they were deleted.
- The following columns are the relevant columns: director_name, duration, gross, genres, movie_title, language, country, budget, title_year, imdb_score.

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

Insights into Task A

- Genre Analysis was done by first converting the genres column into multiple columns of different genres. This was done using text to column with delimiter as ‘ | ’. This allowed us to separate each genre.
- Next, every unique genre was taken and using the COUNTIF function the number of movies for each genre were counted.
- This was followed by using the AVERAGE, MODE, MEDIAN, MAX, MIN, VAR and STDEV functions to calculate the descriptive statistics.
- I then used a column chart to show the most common movie genres.
- The most common genres were Drama, Comedy, Thriller, Action, Romance.
- The genres that weren’t that common were Documentary, Western and Film Noir

- Formulas used:

i. =COUNTIF($A$2:$G$3790,M6)

ii. =AVERAGE(IF(ISNUMBER(SEARCH(M6, $A$2:$G$3724)), $H$2:$H$3724))

iii. =MEDIAN(IF($A$2:$G$3790=M6,$H$2:$H$3790))

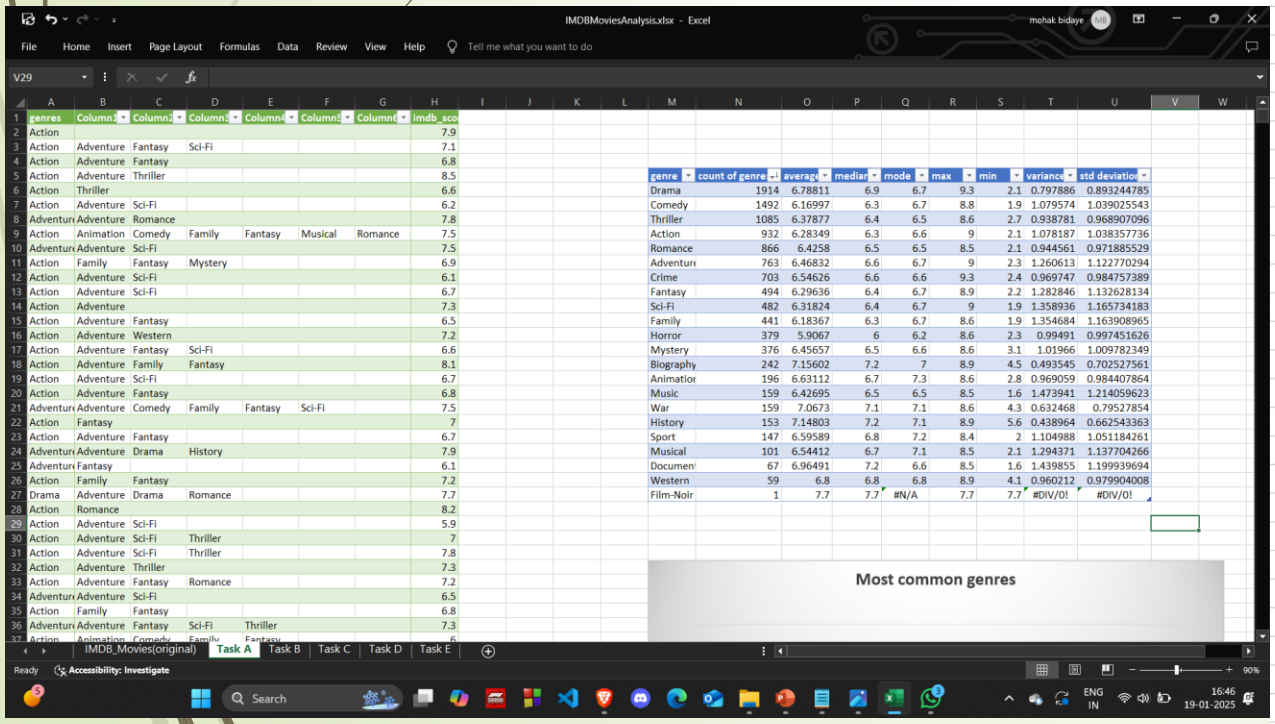
iv. =MODE(IF($A$2:$G$3790=M6,$H$2:$H$3790))

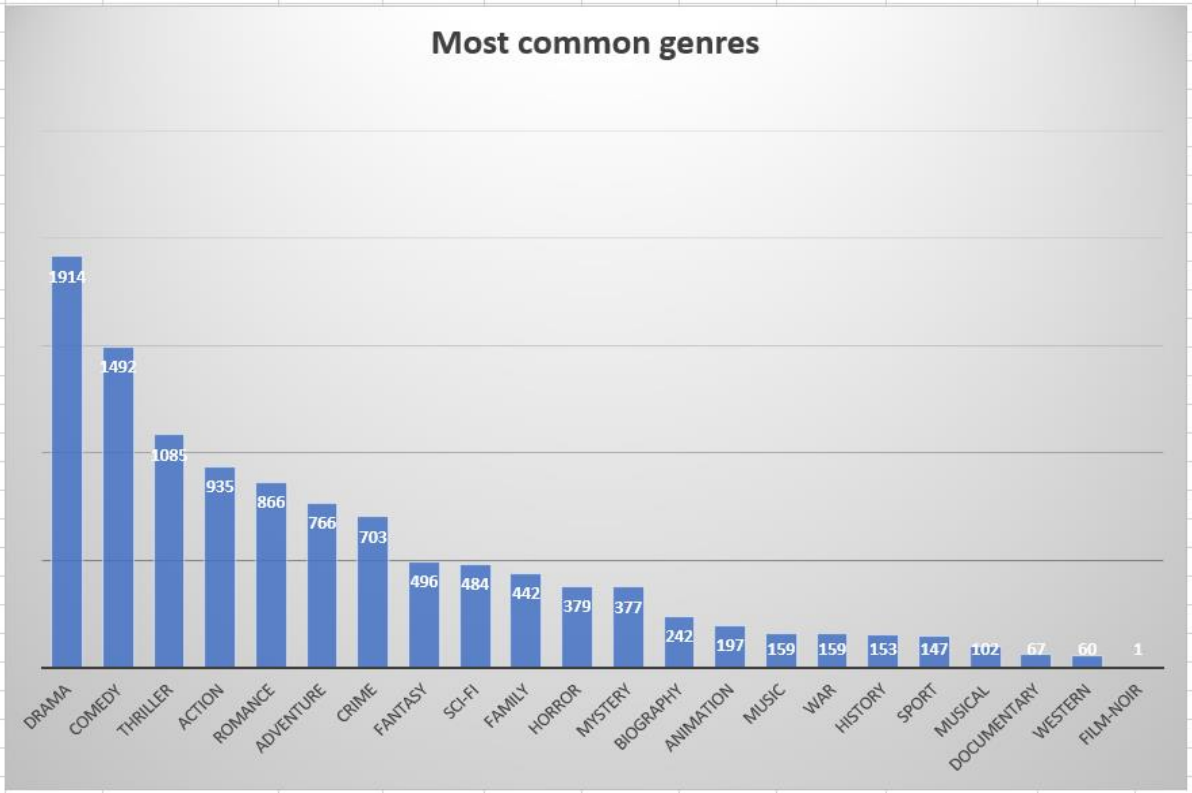
v. =MAX(IF($A$2:$G$3790=M6,$H$2:$H$3790))

vi. =MIN(IF($A$2:$G$3790=M6,$H$2:$H$3790))

vii. =VAR(IF($A$2:$G$3790=M6,$H$2:$H$3790))

viii. =STDEV(IF($A$2:$G$3790=M6,$H$2:$H$3790))





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.

Insights into Task B

- This task was executed by applying the functions average, median and stdev on the duration and imdb_score columns.
- A scatter plot was used to visualize the relationship between the movie duration and imdb score
- From the visualization I understood that the upward slope of the trendline indicates that the movies with a longer duration tend to have a slightly higher imdb rating.
- The scatter plot also suggests that movies within the duration of 90-150 minutes form a dense cluster indicating that most movies fall within this range. These movies receive a mid range imdb score.

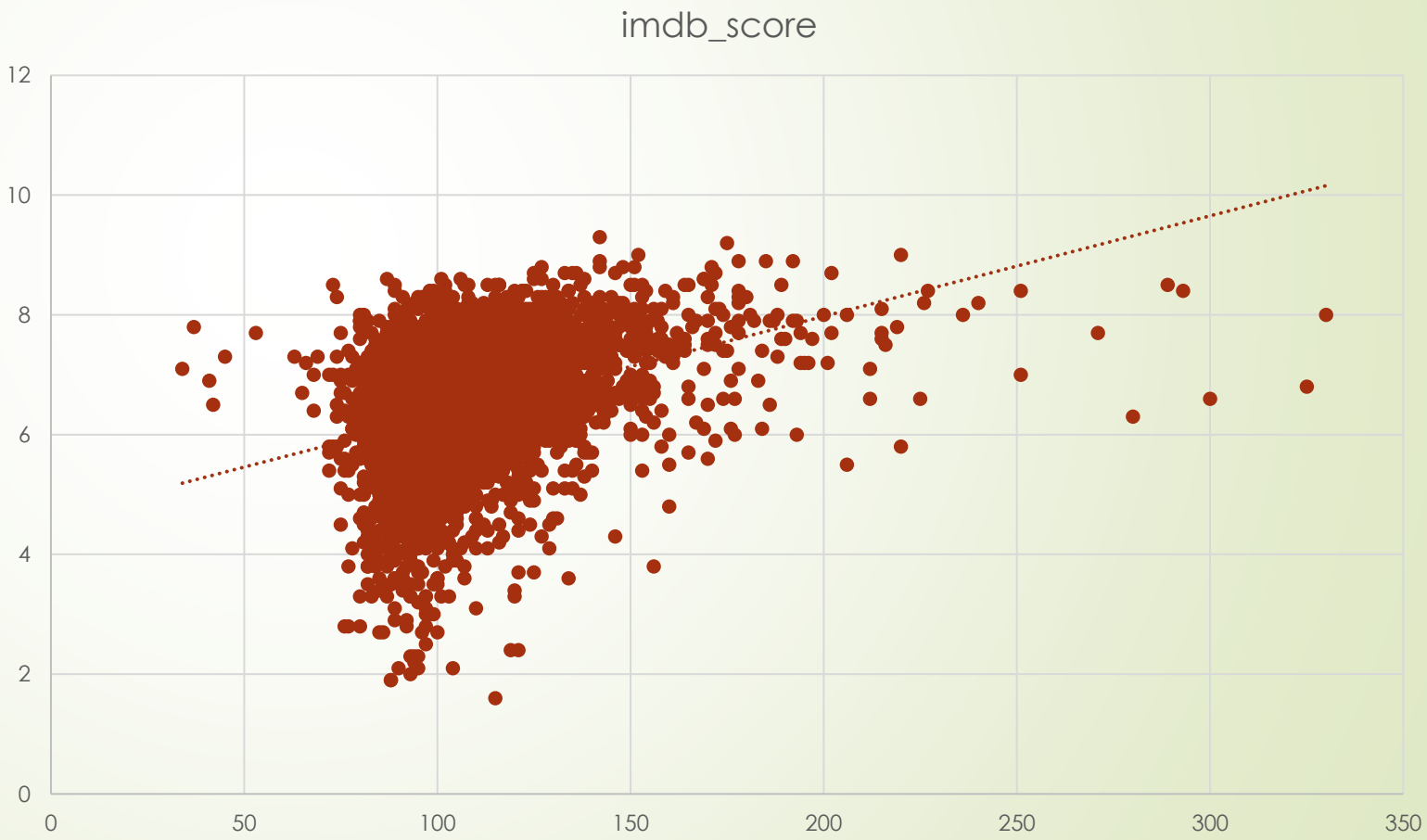
- The formulas I used were:

i. =AVERAGE(A2:A3790)

ii. =MEDIAN(A2:A3790)

iii. =STDEV(A2:A3790)

mean	median	std deviation
109.8029	105	22.75721064



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.

Insights into Task C

- I used the columns movie_title, language and imdb_score.
- By removing the duplicates from the language column I got a list of all the unique languages.
- Next I applied formulas like COUNTIF, AVERAGEIF, MEDIAN, STDEV to get the descriptive statistics.
- I learnt that English was the most popular language for movies. It had a count of 3609. The average imdb score for English movies is 6.42. This is lower than many other languages and a reason for that could be due to the wide variety of movies differing in quality.
- English is followed by French and Spanish as the second and third most common languages.
- The languages having smaller sample sizes indicate that the few movies in these languages tend to receive high ratings. This could be because of the niche and high-quality films.
- Italian and Hindi movies have a broader range of ratings. this shows that their movies can be either of good quality or poor quality.
- Languages like Danish (8.1), Portuguese (8), and Persian (8.4) have high median ratings, suggesting that even the lower-rated movies in these categories are well-received
- Languages with only one movie have no median hence showing an error and no standard deviation.

- Formulas used:

i. =COUNTIF($A$2:$B$3790,E7)

ii. =AVERAGEIF($B$2:$C$3790,E7,$C$2:$C$3790)

iii. =MEDIAN(IF($B$2:$C$3790=E7,$C$2:$C$3790))

iv. =STDEV(IF($B$2:$C$3790=E7,$C$2:$C$3790))

Languages	Count	Average IMDB Score	Mediar	Std Dev
English	3609	6.421030756	6.5	1.052538
French	37	7.286486486	7.2	0.561329
Spanish	26	7.05	7.15	0.826196
Mandarin	14	7.021428571	7.25	0.765786
German	13	7.692307692	7.7	0.640913
Japanese	12	7.625	7.8	0.899621
Hindi	10	6.76	7.05	1.111755
Cantonese	8	7.2375	7.3	0.440576
Italian	7	7.185714286	7	1.155319
Korean	5	7.7	7.7	0.570088
Portuguese	5	7.76	8	0.978775
Norwegian	4	7.15	7.3	0.574456
Danish	3	7.9	8.1	0.52915
Dutch	3	7.566666667	7.8	0.404145
Hebrew	3	7.5	7.3	0.43589
Persian	3	8.133333333	8.4	0.550757
Thai	3	6.633333333	6.6	0.450925
Aboriginal	2	6.95	6.95	0.777817
Dari	2	7.5	7.5	0.141421
Indonesian	2	7.9	7.9	0.424264
Arabic	1	7.2	7.2	#DIV/0!
Aramaic	1	7.1	7.1	#DIV/0!
Bosnian	1	4.3	4.3	#DIV/0!
Czech	1	7.4	7.4	#DIV/0!
Dzongkha	1	7.5	7.5	#DIV/0!
Filipino	1	6.7	6.7	#DIV/0!
Hungarian	1	7.1	7.1	#DIV/0!
Icelandic	1	6.9	6.9	#DIV/0!
Kazakh	1	6	6	#DIV/0!
Maya	1	7.8	7.8	#DIV/0!
Mongolian	1	7.3	7.3	#DIV/0!
None	1	8.5	8.5	#DIV/0!
Romanian	1	7.9	7.9	#DIV/0!
Russian	1	6.5	6.5	#DIV/0!
Swedish	1	7.6	7.6	#DIV/0!
Telugu	1	8.4	8.4	#DIV/0!
Vietnamese	1	7.4	7.4	#DIV/0!
Zulu	1	7.3	7.3	#DIV/0!

Task D:
Director Analysis: Influence of directors on movie ratings.

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

Insights into Task D:

- The directors column was taken and duplicates were removed to get a new column of all the unique director names.
- Next, I calculated the average of the imdb score for each director by using the AVERAGEIF formula.
- Using a pivot table I've displayed the average imdb score of the top 10 directors.
- The percentile function gave a value of 7.5 which indicated that the directors with a score above 7.5 are the top directors.
- I then labelled the top directors with the IF function.
- A pie chart was made to represent the top 10 directors. This was done by using the pivot table.

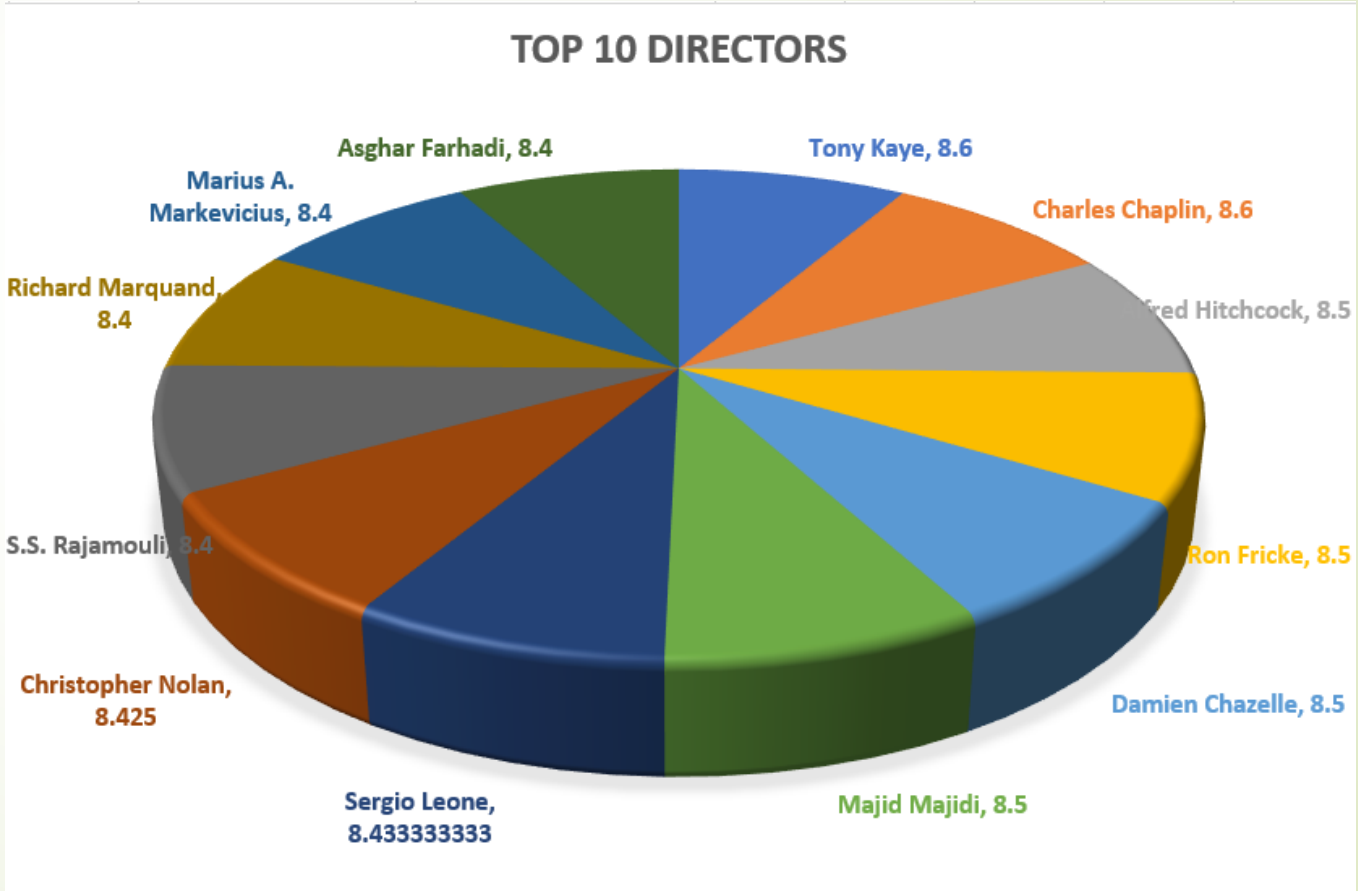
- Formulas used:

i. =AVERAGEIF($A:$A,D2,$B:$B)

ii. =PERCENTILE(E:E,0.9)

iii. =IF(E2>7.5,"Top","")

percentile
7.5



TASK E:

Budget Analysis: 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.

Insights into Task E:

- Profit was calculated by subtracting gross from the budget.
- The top 10 movies with the most profit are displayed using a pivot table.
- CORREL function was used to calculate the correlation coefficient between gross and budget.
- A correlation of 0.09664 suggests that there is a very weak relationship between a movies budget and its gross earnings.
- This made me understand that budget alone doesn't drive success and that gross earnings depend on various factors like marketing, audience appeal, star power and release timing which may overshadow the impact of the budget.
- The movie with the most profit was Avatar. It had a profit of $52,35,05,847.

- Formulas used:

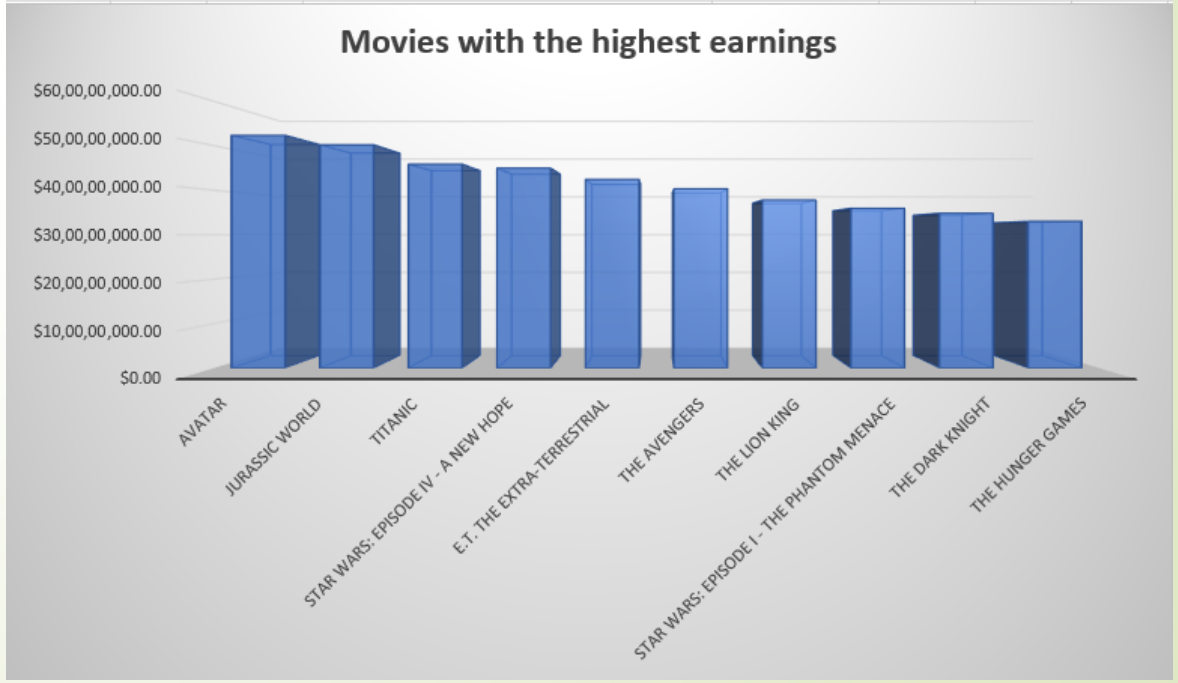
i. Gross-budget (C2-B2)

ii. =CORREL(B:B,C:C)

iii. =MAX(D:D)

movie_title	budget	gross	profit	correlat	maximum
Avatar	$23,70,00,000	$76,05,05,847	$52,35,05,847	0.096648	$52,35,05,847
Jurassic World	$15,00,00,000	$65,21,77,271	$50,21,77,271		
Titanic	$20,00,00,000	$65,86,72,302	$45,86,72,302		
Star Wars: Episode IV - A New Hope	$1,10,00,000	$46,09,35,665	$44,99,35,665		
E.T. the Extra-Terrestrial	$1,05,00,000	$43,49,49,459	$42,44,49,459		
The Avengers	$22,00,00,000	$62,32,79,547	$40,32,79,547		
The Lion King	$4,50,00,000	$42,27,83,777	$37,77,83,777		
Star Wars: Episode I - The Phantom Menace	$11,50,00,000	$47,45,44,677	$35,95,44,677		
The Dark Knight	$18,50,00,000	$53,33,16,061	$34,83,16,061		
The Hunger Games	$7,80,00,000	$40,79,99,255	$32,99,99,255		
Deadpool	$5,80,00,000	$36,30,24,263	$30,50,24,263		
The Hunger Games: Catching Fire	$13,00,00,000	$42,46,45,577	$29,46,45,577		
Jurassic Park	$6,30,00,000	$35,67,84,000	$29,37,84,000		
Despicable Me 2	$7,60,00,000	$36,80,49,635	$29,20,49,635		
American Sniper	$5,88,00,000	$35,01,23,553	$29,13,23,553		
Finding Nemo	$9,40,00,000	$38,08,38,870	$28,68,38,870		
Shrek 2	$15,00,00,000	$43,64,71,036	$28,64,71,036		
The Lord of the Rings: The Return of the King	$9,40,00,000	$37,70,19,252	$28,30,19,252		
Star Wars: Episode VI - Return of the Jedi	$3,25,00,000	$30,91,25,409	$27,66,25,409		
Forrest Gump	$5,50,00,000	$32,96,91,196	$27,46,91,196		

Row Labels	Sum of profit
Avatar	$52,35,05,847.00
Jurassic World	$50,21,77,271.00
Titanic	$45,86,72,302.00
Star Wars: Episode IV - A New Hope	$44,99,35,665.00
E.T. the Extra-Terrestrial	$42,44,49,459.00
The Avengers	$40,32,79,547.00
The Lion King	$37,77,83,777.00
Star Wars: Episode I - The Phantom Menace	$35,95,44,677.00
The Dark Knight	$34,83,16,061.00
The Hunger Games	$32,99,99,255.00
Grand Total	**4177663861**



CONCLUSION

- Through this project I was able to gain hands-on learning experience and improve my skills in data analysis, statistical methods and visualization tools. My knowledge and understanding of how data can tell a story and provide meaningful insights has also deepened.

Links

- Excel Sheet link:

https://docs.google.com/spreadsheets/d/1SWtoqWblJ5XywaWvrovBgeiBqEgtynuJ/edit?usp=drive_link&ouid=109524556463170667809&rtpof=true&sd=true