IMDB MOVIE ANALYSIS

Video Link -

https://drive.google.com/file/d/1v6he_JFuQzOHaofSsnlWyy1mD8R5XxUy/view?usp=sharing

Project Description:

In this project we are provided with a dataset of IMDB Movies. We need to investigate the factors that are responsible for the success of a movie on IMDB. Success maybe high IMDB ratings. This will make an impact on movies producers and directors to understand what makes a movie successful.

Approach:

- Downloaded the data.

- Tried to understand the data.

- Cleaned the data.

- Performed certain formulas.

- Found the answers.

Cleaning of Data:

- Removing the rows having blank cells or null cells.
- After reading and understanding the tasks, I found that only these columns are necessary for the analysis:-
  genres, imdb_score, duration, language, director_name, budget, gross
  rest all columns were of no use so I removed them all.
- Removed the duplicate rows.
  Cleaned data file -
  https://drive.google.com/file/d/1_XyY39_uZUgweDLbg3SIPlQn__9j7dXp/view?usp=sharing

Tech-Stack used:

Microsoft Excel

It is a spreadsheet software which is used to perform certain operations on spreadsheets. It also allows us to visualize the data using bar graphs etc.

Insights:

(A) Movie Genre Analysis

In this task I had to find the most common genres of movies in the dataset and then calculating the mean, median, mode, range, variance, standard deviation of IMDB scores for each genres.

- First, I separated the data from multiple genres to single genres.
- Then, I created a new table of Genre and number of Movies.
- Using countifs, calculated number of movies for each genre.
- The required output looks like this.

Genre	Movies
Drama	1846
Comedy	1443
Thriller	1071
Action	923
Romance	838
Adventure	751
Crime	692
Fantasy	486
Sci-Fi	479
Family	431
Horror	374
Mystery	371
Biography	237
Animation	194
Music	149
War	149
History	146
Sport	143
Musical	95
Western	57
Documentary	45
Film-Noir	1
Total	10921

Then I used Average, Median, Mode, Max, Min, Var and STDEV functions to find descriptive analysis.

Drama, Comedy, Thriller were the genres having highest number of movies but the mean was around 6.4 for them which not that high.

When we talk about mean IMDB score, War, history, Biography were the movies with mean IMDB score above 7. We are not including the data of Film-Noir as we have record of only one movie for this category. When we compared standard deviation with them then for History Genre the standard deviation was lowest so we can say that most the History movies IMDB score lies around 7.15.

But if we talk about more data of movies which gives us more accurate analysis then we found that Drama movies scores around 6.8 having more than 600-700 movies. And the maximum drama IMDB score was the highest among all.

Genre	Movies	Mean	Median	Mode	Min	Max	Variance	Standard Deviation
Drama	1846	6.795666	6.9	6.7	2.1	9.3	0.7934338	0.890749001
Comedy	1443	6.186972	6.3	6.7	1.9	8.8	1.0707594	1.034775052
Thriller	1071	6.372923	6.4	6.5	2.7	9	0.9424437	0.970795392
Action	923	6.284507	6.3	6.6	2.1	9	1.0808985	1.039662703
Romance	838	6.432458	6.5	6.5	2.1	8.5	0.9193753	0.958840608
Adventure	751	6.448336	6.6	6.6	2.3	8.9	1.2598339	1.122423223
Crime	692	6.545231	6.6	6.6	2.4	9.3	0.9720612	0.985931621
Fantasy	486	6.28107	6.4	6.7	2.2	8.9	1.2958059	1.138334689
Sci-Fi	479	6.33048	6.4	7	1.9	8.8	1.3673117	1.16932104
Family	431	6.2058	6.3	6.1	1.9	8.6	1.3711988	1.170981996
Horror	374	5.9	5.9	5.9	2.3	8.6	0.9818231	0.990869848
Mystery	371	6.477358	6.5	6.6	3.1	8.6	1.0096481	1.004812489
Biography	237	7.156118	7.2	7	4.5	8.9	0.4811171	0.693626031
Animation	194	6.702577	6.8	7	2.8	8.6	0.9892679	0.994619492
Music	149	6.336913	6.5	6.5	1.6	8.5	1.5180201	1.232079597
War	149	7.057047	7.1	7.1	4.3	8.6	0.6432777	0.802045951
History	146	7.15274	7.2	7.7	5.5	8.9	0.4443028	0.666560415
Sport	143	6.586713	6.8	7.2	2	8.3	1.100456	1.049026226
Musical	95	6.588421	6.7	6.2	2.1	8.5	1.2201837	1.104619233
Western	57	6.812281	6.8	6.8	4.7	8.9	0.8857393	0.941137263
Document	45	6.988889	7.4	7.6	1.6	8.5	1.9173737	1.384692651
Film-Noir	1	7.7	7.7	#N/A				
Total	10921							

File link –

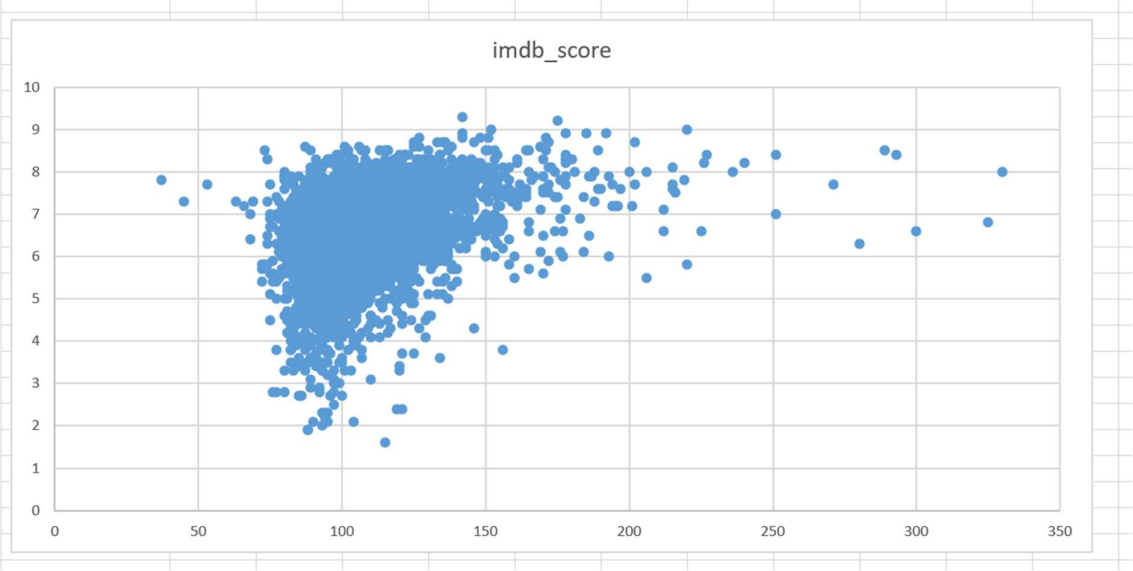
https://docs.google.com/spreadsheets/d/1XU1JRZOx6x1vvJJzRnPiXSKcLZOpj7Wc/edit?usp=sharing&ouid=112124140114283693349&rtpof=true&sd=true

(B) Movie Duration Analysis

In this task we need only two columns, i.e Movie duration and IMDB score. I calculated the Descriptive Analysis such as Mean, Median, Standard Deviation by using the Excel functions Average, Median, STDEV.

Mean	Median	Standard Deviation
110.1524	106	22.69031468

To plot the scatter plot, I selected the data then in Insert I selected Scatter X Y and the plotted the selected graph.



The average length of movies was 110 minutes. After watching the Scatter plot of movie duration and IMDB scores and ignoring the outliers we can find that movies between 110 to 150 minutes are getting average IMDB score more than 6 and the highest IMDB score also lies in this category only. The number of movies getting score below 4 is highest for the duration between 50-100 minutes.

File link – https://docs.google.com/spreadsheets/d/10zbdru4EeWyALEEOZGx-G9K7HC4jo8fx/edit?usp=sharing&ouid=112124140114283693349&rtpof=true&sd=true

(C) Language Analysis:

I calculated the number of movies with language using countif function and descriptive analysis using Average, median and stdev functions.

Language	Movies	Mean	Median	Standard Deviation
Aboriginal	2	6.95	6.95	0.777817459
Arabic	1	7.2	7.2	#DIV/0!
Aramaic	1	7.1	7.1	#DIV/0!
Bosnian	1	4.3	4.3	#DIV/0!
Cantonese	7	7.34286	7.3	0.350509833
Czech	1	7.4	7.4	#DIV/0!
Danish	3	7.9	8.1	0.529150262
Dari	2	7.5	7.5	0.141421356
Dutch	3	7.56667	7.8	0.404145188
English	3498	6.42441	6.5	1.052134702
Filipino	1	6.7	6.7	#DIV/0!
French	34	7.35588	7.3	0.519435111
German	10	7.77	7.8	0.711883261
Hebrew	1	8	8	#DIV/0!
Hindi	5	7.22	7.4	0.801249025
Hungarian	1	7.1	7.1	#DIV/0!
Indonesian	2	7.9	7.9	0.424264069
Italian	7	7.18571	7	1.155318962
Japanese	10	7.66	8	0.990173947
Kazakh	1	6	6	#DIV/0!
Korean	5	7.7	7.7	0.570087713
Mandarin	14	7.02143	7.25	0.765786244
Maya	1	7.8	7.8	#DIV/0!
Mongolian	1	7.3	7.3	#DIV/0!
None	1	8.5	8.5	#DIV/0!
Norwegian	4	7.15	7.3	0.574456265
Persian	3	8.13333	8.4	0.550757055
Portuguese	5	7.76	8	0.978774744
Romanian	1	7.9	7.9	#DIV/0!
Russian	1	6.5	6.5	#DIV/0!
Spanish	23	7.08261	7.2	0.860577065
Thai	3	6.63333	6.6	0.450924975
Vietnamese	1	7.4	7.4	#DIV/0!
Zulu	1	7.3	7.3	#DIV/0!
Vietnamese	1	7.4	7.4	#DIV/0!
Zulu	1	7.3	7.3	#DIV/0!

The maximum movies were made in English language. French and Spanish were having 34 and 23 movies respectively. So only this much data was useful. The mean for these languages was higher but we can't compare that data with the data of English language as it is having too many movies.

File link -

https://docs.google.com/spreadsheets/d/1_2LDMC8fOrZqBzERm2bYv_OZ8-AGV7g6/edit?usp=sharing&ouid=112124140114283693349&rtpof=true&sd=true

(D) Director Analysis

To do this task, I selected two columns i.e Director name and imdb score then using avg if function I found the avg IMDB score of every director then to find the percentile and rank, I selected data then data analysis then select rank and percentile option from it to find the rank and percentile for the whole data.

The top 10 directors are:-

Director_name	AVG IMDB Score	Rank	Percentile
Akira Kurosawa	8.7	1	100.00%
Tony Kaye	8.6	2	99.80%
Charles Chaplin	8.6	2	99.80%
Alfred Hitchcock	8.5	4	99.60%
Ron Fricke	8.5	4	99.60%
Damien Chazelle	8.5	4	99.60%
Majid Majidi	8.5	4	99.60%
Sergio Leone	8.433333333	8	99.50%
Christopher Nolan	8.425	9	99.50%
Richard Marquand	8.4	10	99.30%

From this analysis we were able to see the top 10 directors with highest average of IMDB score. The biggest insight from this analysis we found is that the average of those top directors was too high around 8.5. So we can clearly state that the success of movie is highly dependent on the director of that movie.

File link -

https://docs.google.com/spreadsheets/d/19RClBKO6qFboV3nj8G7uHTZx14IQJVDU/edit?usp=sharing&ouid=112124140114283693349&rtpof=true&sd=true

(E) Budget Analysis

To do this task, we need three columns that are movie_title, budget, gross and with budget and gross I created a new column that is profit margin by subtracting budget from gross.

Now, to find the corelation between budget and gross I used the CORREL formula with budget and gross i.e =correl(B:B,C:C). The correlation found was 0.093 or we can say 0.01 which is very low. So, we can say that movies gross margin depends very low on its budget.

Corelation
0.093445

I sorted the profit margin from largest to smallest through which I found the movies with highest profit margins.

movie_title	budget	gross	Profit Margin
Avatar	237000000	760505847	523505847
Jurassic World	150000000	652177271	502177271
Titanic	200000000	658672302	458672302
Star Wars: Episode IV - A New Hope	11000000	460935665	449935665
E.T. the Extra-Terrestrial	10500000	434949459	424449459
The Avengers	220000000	623279547	403279547
The Lion King	45000000	422783777	377783777
Star Wars: Episode I - The Phantom Menace	115000000	474544677	359544677
The Dark Knight	185000000	533316061	348316061
The Hunger Games	78000000	407999255	329999255
Deadpool	58000000	363024263	305024263
The Hunger Games: Catching Fire	130000000	424645577	294645577
Jurassic Park	63000000	356784000	293784000
Despicable Me 2	76000000	368049635	292049635
American Sniper	58800000	350123553	291323553
Finding Nemo	94000000	380838870	286838870
Shrek 2	150000000	436471036	286471036
The Lord of the Rings: The Return of the King	94000000	377019252	283019252
Star Wars: Episode VI - Return of the Jedi	32500000	309125409	276625409
Forrest Gump	55000000	329691196	274691196
Star Wars: Episode V - The Empire Strikes Back	18000000	290158751	272158751
Home Alone	18000000	285761243	267761243
Star Wars: Episode III - Revenge of the Sith	113000000	380262555	267262555
Spider-Man	139000000	403706375	264706375
Minions	74000000	336029560	262029560
The Sixth Sense	40000000	293501675	253501675
Jaws	8000000	260000000	252000000
Frozen	150000000	400736600	250736600
The Secret Life of Pets	75000000	323505540	248505540
The Twilight Saga: New Moon	50000000	296623634	246623634

It is Avatar on top.

From this analysis we found that the correlation between budget and gross is 0.093 which is too low. This means that the earning of movie depends too low on its budget as low budget can increase the profit margin and vice-versa.

File link - https://docs.google.com/spreadsheets/d/18feNd0mP6Y13NK4VrNLIg0fqkbx_zp8T/edit?usp=sharing&ouid=112124140114283693349&rtpof=true&sd=true

Result:

(A) – Movie Genre Analysis

- Drama movies getting highest Mean, Max IMDB scores with having more than 1800 movies. It is extremely beneficial to make a movie on Drama genre.
- History movies can do good. Creators can also work on history movies to get an average IMDB score as the standard deviation is low for this genre with high mean IMDB score.

(B) – Movie Duration Analysis

- The best duration for the movies will be around 130-150 minutes with low risk of low IMDB score and high chances of high IMDB score.
- Creating movies with less than 100 minutes of watch time is highly risky as we can see very low scores in that category. This category must be avoided.

(C) - Language Analysis

- To make the movies to reach to maximum audience, the language must be English.
- The mean of French and Spanish were also good. If we are going to make movie on some other language then we should try them.

(D) Director Analysis

- The top 10 directors were Akira Kurosawa, Tony Kaye, Charles Chaplin, Alfred Hitchcock, Ron Fricke, Damien Chazelle, Majid Majidi, Sergio Leone, Chritopher Nolan and Richard Marquand with Akira Kurosawa as the top director.

(E) Budget Analysis

- Instead of focusing on budget we should focus on better directors and actors for that movie. It may increase the chances of success.

Video Link -

https://drive.google.com/file/d/1v6he_JFuQzOHaofSsnlWyy1mD8R5XxUy/view?usp=sharing