IMDB Movie Analysis

Project Description: This project is about IMDB Movie Analysis. I was provided with the dataset having various columns of different IMDB Movies. I was required to provide a detailed report for the below data record mentioning the answers of the questions that follows:

1. **Movie Genre Analysis:** Analyze the distribution of movie genres and their impact on the IMDB score.
   My 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.

2. **Movie Duration Analysis:** Analyze the distribution of movie durations and its impact on the IMDB score.
   My Task: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

3. **Language Analysis:** Situation: Examine the distribution of movies based on their language.
   My Task: Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.

4. **Director Analysis:** Influence of directors on movie ratings.
   My Task: Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

5. **Budget Analysis:** Explore the relationship between movie budgets and their financial success.
   My Task: Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

Approach: First I gone through dataset to know all the columns present in the table. Then I saw all the questions and thought of functions which could be used to answer each question. After that I applied those functions and found the answer to each question and plotted the graph wherever was required. I started with cleaning the data set using following steps:

- First, I deleted all the unwanted columns which were: color, director_facebook_likes, actor_3_facebook_likes, actor_1_facebook_likes, cast_total_facebook_likes, facenumber_in_poster, plot_keywords, actor_2_facebook_likes, aspect_ratio.
- After that I removed all the rows with blank cells. For this first, I selected whole table. After that, I selected go to special option inside find & select menu. Inside the menu I selected blanks option and all the blank cells were selected. Then I simply deleted the rows which consists those blank cells.
- Lastly, I removed all the duplicate rows.

Link for cleaned data :-

https://docs.google.com/spreadsheets/d/1GHW0M2VqPRlTDKSleDtSgmGEIAcq7bvh/edit?usp=sharing&ouid=106942457558004201317&rtpof=true&sd=true

Tech-Stack Used: The software used for the project is Microsoft Excel 365. It is used to run the functions and get answers of each question. It is also used to plot the graphs.

Insights:

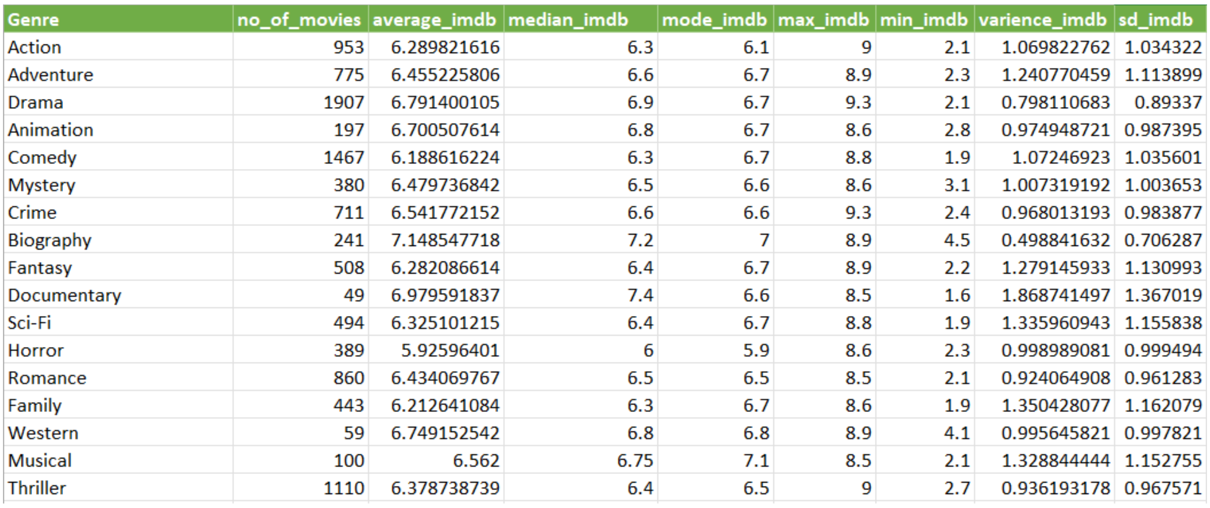
1. Movie Genre Analysis:

   Function:-

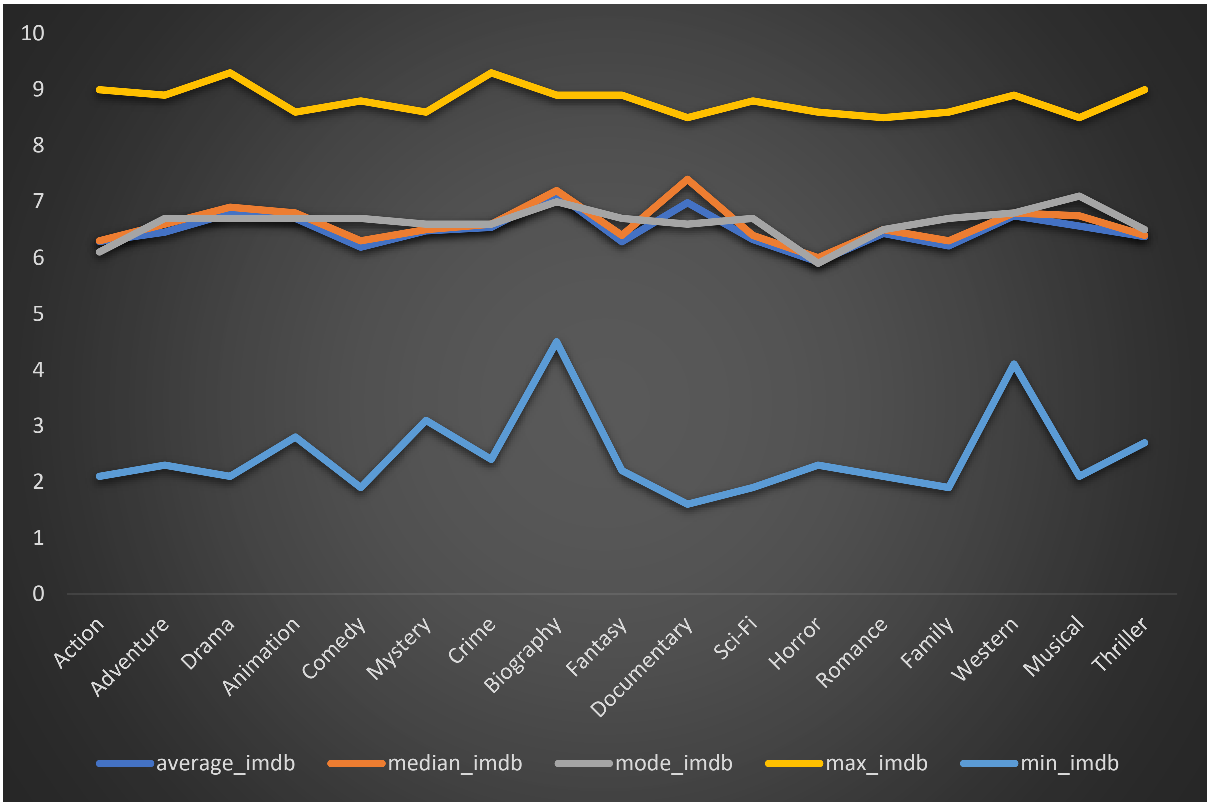
```
=IF(ISNUMBER(SEARCH("|",Table_IMDB_Movies[@genres])),LEFT(Table_IMDB_Movies[@genres],SEARCH("|",Table_IMDB_Movies[@genres])-1),Table_IMDB_Movies[@genres])
=UNIQUE(V2:V3768)
=COUNTIF(F$2:F$3768,"*" & X2 & "*")
=AVERAGEIF(F$2:F$3768,"*" & X2 & "*",R$2:R$3768)
=MEDIAN(IF(ISNUMBER(SEARCH("*" & X2 & "*",F$2:F$3768)),R$2:R$3768))
=MODE(IF(ISNUMBER(SEARCH("*" & X2 & "*",F$2:F$3768)),R$2:R$3768))
=MAXIFS(R$2:R$3768,F$2:F$3768,"*" & X2 & "*")
=MINIFS(R$2:R$3768,F$2:F$3768,"*" & X2 & "*")
=VAR(IF(ISNUMBER(SEARCH("*" & X2 & "*",F$2:F$3768)),R$2:R$3768))
=STDEV(IF(ISNUMBER(SEARCH("*" & X2 & "*",F$2:F$3768)),R$2:R$3768))
```

Output:-

https://docs.google.com/spreadsheets/d/1SwIZECZBaX83YIgE7s_fAiMSR1AVUm2f/edit?usp=sharing&ouid=106942457558004201317&rtpof=true&sd=true



Genre	no_of_movies	average_imdb	median_imdb	mode_imdb	max_imdb	min_imdb	varience_imdb	sd_imdb
Action	953	6.289821616	6.3	6.1	9	2.1	1.069822762	1.034322
Adventure	775	6.455225806	6.6	6.7	8.9	2.3	1.240770459	1.113899
Drama	1907	6.791400105	6.9	6.7	9.3	2.1	0.798110683	0.89337
Animation	197	6.700507614	6.8	6.7	8.6	2.8	0.974948721	0.987395
Comedy	1467	6.188616224	6.3	6.7	8.8	1.9	1.07246923	1.035601
Mystery	380	6.479736842	6.5	6.6	8.6	3.1	1.007319192	1.003653
Crime	711	6.541772152	6.6	6.6	9.3	2.4	0.968013193	0.983877
Biography	241	7.148547718	7.2	7	8.9	4.5	0.498841632	0.706287
Fantasy	508	6.282086614	6.4	6.7	8.9	2.2	1.279145933	1.130993
Documentary	49	6.979591837	7.4	6.6	8.5	1.6	1.868741497	1.367019
Sci-Fi	494	6.325101215	6.4	6.7	8.8	1.9	1.335960943	1.155838
Horror	389	5.92596401	6	5.9	8.6	2.3	0.998989081	0.999494
Romance	860	6.434069767	6.5	6.5	8.5	2.1	0.924064908	0.961283
Family	443	6.212641084	6.3	6.7	8.6	1.9	1.350428077	1.162079
Western	59	6.749152542	6.8	6.8	8.9	4.1	0.995645821	0.997821
Musical	100	6.562	6.75	7.1	8.5	2.1	1.328844444	1.152755
Thriller	1110	6.378738739	6.4	6.5	9	2.7	0.936193178	0.967571



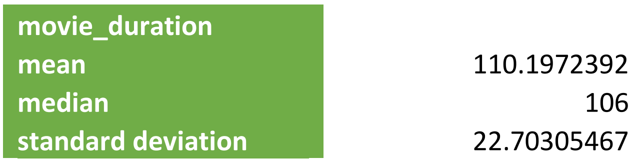
2. Movie Duration Analysis:

Function:-

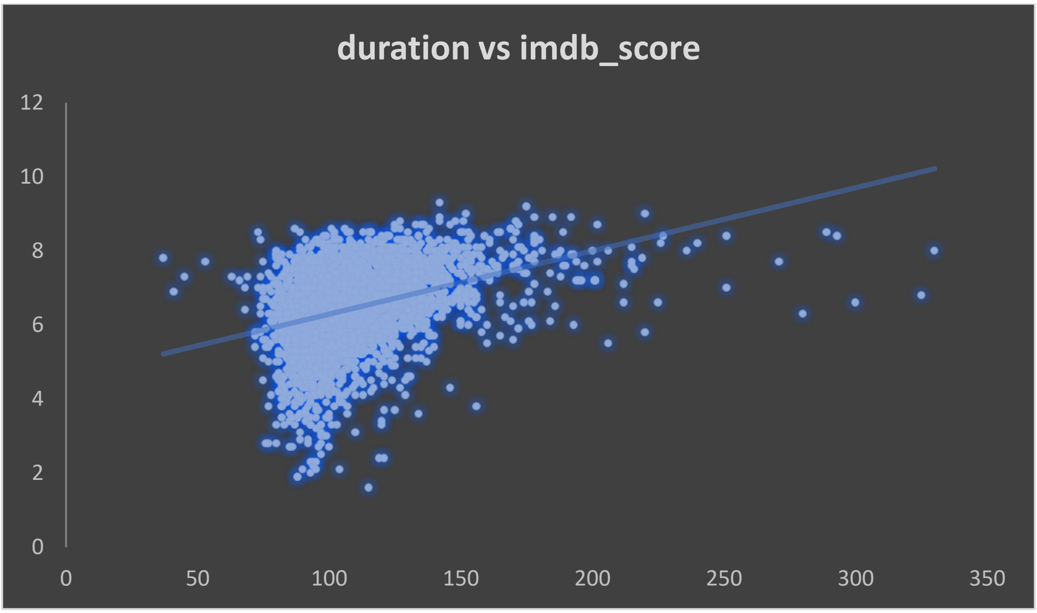
=AVERAGE(C2:C3768)
=MEDIAN(C2:C3768)
=STDEVA(C2:C3768)

Output:-

https://docs.google.com/spreadsheets/d/1a3f0MsCE5PHjF2Dn7WTYn9tZekDIO7F-/edit?usp=sharing&ouid=106942457558004201317&rtpof=true&sd=true



movie_duration	
mean	110.1972392
median	106
standard deviation	22.70305467



3. Language Analysis:

   Function:-

   =UNIQUE(M2:M3768)
   =COUNTIF(M$2:M$3768,U2)
   =AVERAGEIF(M$2:M$3768,U2,R$2:R$3768)
   =MEDIAN(IF(M$2:M$3768=U2,R$2:R$3768))
   =STDEV(IF(M$2:M$3768=U2,R$2:R$3768))

   Output:-

   https://docs.google.com/spreadsheets/d/1cmXAIl9iLqLw3nV1WU8y3tSw62Kij-Q8/edit?usp=sharing&ouid=106942457558004201317&rtpof=true&sd=true

language	no_of_movies	mean_imdb	median_imdb	stand_imdb
English	3593	6.42599499	6.5	1.049712558
Mandarin	14	7.021428571	7.25	0.765786244
Aboriginal	2	6.95	6.95	0.777817459
Spanish	23	7.082608696	7.2	0.860577065
French	36	7.297222222	7.25	0.565425812
Filipino	1	6.7	6.7	#DIV/0!

Maya	1	7.8	7.8	#DIV/0!
Kazakh	1	6	6	#DIV/0!
Telugu	1	8.4	8.4	#DIV/0!
Cantonese	8	7.2375	7.3	0.440575922
Japanese	12	7.625	7.8	0.899621132
Aramaic	1	7.1	7.1	#DIV/0!
Italian	7	7.185714286	7	1.155318962
Dutch	3	7.566666667	7.8	0.404145188
Dari	2	7.5	7.5	0.141421356
German	13	7.692307692	7.7	0.640912811
Mongolian	1	7.3	7.3	#DIV/0!
Thai	3	6.633333333	6.6	0.450924975
Bosnian	1	4.3	4.3	#DIV/0!
Korean	5	7.7	7.7	0.570087713
Hungarian	1	7.1	7.1	#DIV/0!
Hindi	8	7.175	7.3	0.761108215
	2	5.3	5.3	0.707106781
Icelandic	1	6.9	6.9	#DIV/0!
Danish	3	7.9	8.1	0.529150262
Portuguese	5	7.76	8	0.978774744
Norwegian	4	7.15	7.3	0.574456265
Czech	1	7.4	7.4	#DIV/0!
Russian	1	6.5	6.5	#DIV/0!
None	1	8.5	8.5	#DIV/0!
Zulu	1	7.3	7.3	#DIV/0!
Hebrew	1	8	8	#DIV/0!
Dzongkha	1	7.5	7.5	#DIV/0!
Arabic	1	7.2	7.2	#DIV/0!
Vietnamese	1	7.4	7.4	#DIV/0!
Indonesian	2	7.9	7.9	0.424264069
Romanian	1	7.9	7.9	#DIV/0!
Persian	3	8.133333333	8.4	0.550757055
Swedish	1	7.6	7.6	#DIV/0!

4. Director Analysis:

Function:-

```
=UNIQUE(A2:A3768)
=AVERAGEIF(A$2:A$3768,V2,R$2:R$3768)
=PERCENTRANK.EXC(W$2:W$1694,W2)
=INDEX(SORTBY(V2:W1694,X2:X1694,-1,V2:V1694,1),SEQUENCE(10),{1,2})
```

```
=AB2-AVERAGE(W$2:W$1694)
```

To compare the scores of top directors to the overall distribution of scores I subtracted average imdb score from imdb score of the director.

Output:-

https://docs.google.com/spreadsheets/d/11wWcvzX2oJnM-TvFBJpJpqaH7SDqdFCW/edit?usp=sharing&ouid=106942457558004201317&rtpof=true&sd=true

Rank	top10director	idmb_score	greater_than _average_score
1	Charles Chaplin	8.6	2.287152896
2	Tony Kaye	8.6	2.287152896
3	Alfred Hitchcock	8.5	2.187152896
4	Damien Chazelle	8.5	2.187152896
5	Majid Majidi	8.5	2.187152896
6	Ron Fricke	8.5	2.187152896
7	Christopher Nolan	8.425	2.112152896
8	Sergio Leone	8.433333333	2.120486229
9	Asghar Farhadi	8.4	2.087152896
10	Richard Marquand	8.4	2.087152896

5. Popular Genres:

Function:-

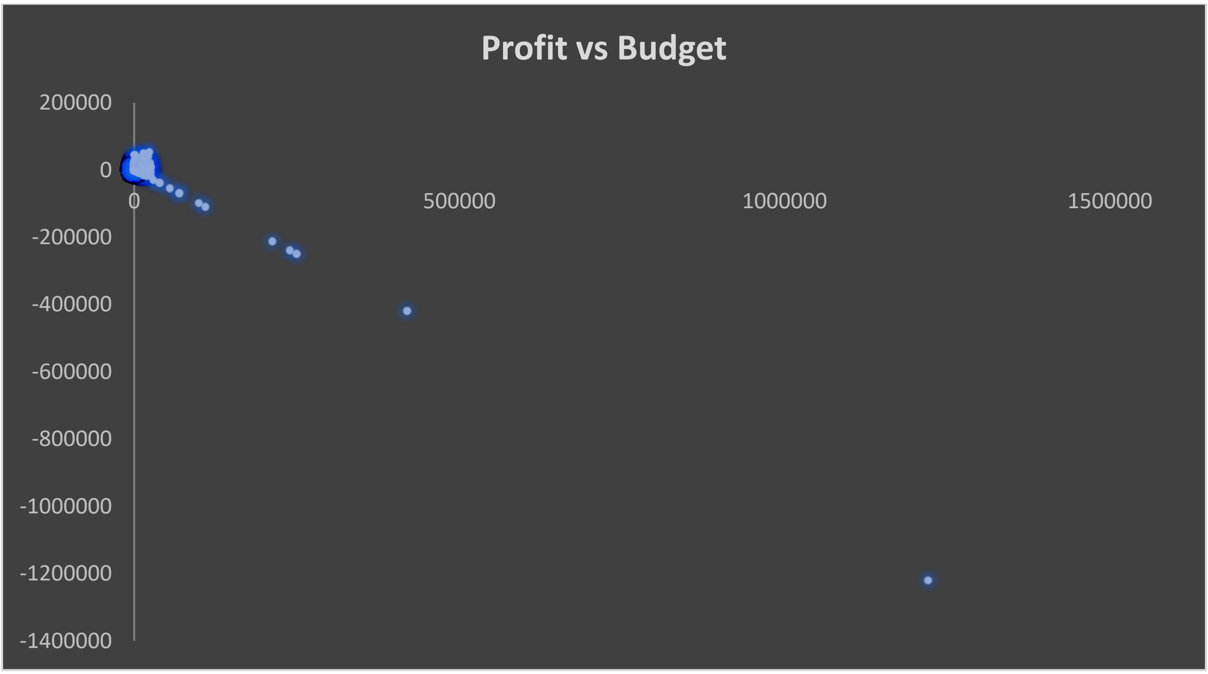
```
=[@gross]-[@budget]
=CORREL(P2:P3768,E2:E3768)
=IF(T2:T3768=MAX(T2:T3768),H2:H3768,"")
=INDEX(SORTBY(H2:H3768,T2:T3768,-1),SEQUENCE(5))
```

Output:-

https://docs.google.com/spreadsheets/d/1IA7UlJ3YgYshWf3XKh6KGA8gbk76AqtE/edit?usp=sharing&ouid=106942457558004201317&rtpof=true&sd=true

Movie with highest profit :- Avatar

movies_with_highest_profit
Avatar
Jurassic World
Titanic
Star Wars: Episode IV - A New Hope
E.T. the Extra-Terrestrial

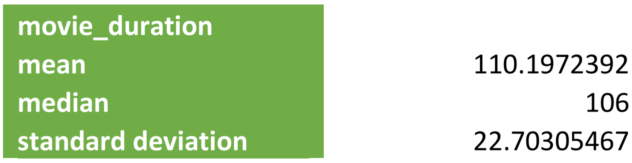


Results:

Cleaning the data: Cleaned data consists of 3768 rows including title.

1. Movie Genre Analysis:
   The impact of genre is almost same on mean, median, mode movie ratings

2. Movie Duration Analysis:



movie_duration	
mean	110.1972392
median	106
standard deviation	22.70305467

3. Language Analysis:
   English is most popular language.

4. Director Analysis:
   Charles Chaplin is the director with highest imdb rating.

5. Budget Analysis:
   Avatar is the movie with highest profit.