Trainity

IMDB Movie Analysis

Final Project – 1

Done by

Gaddam Lakshmi Deepak

Data Analytics Trainee

Description

This project is for analysing the dataset related to IMDB Movies and draw meaningful insights from it. I have done the required analysis to get the solution. By doing this task which helps for movie producers, directors and investors who want to get understanding about making movie successful. For this project, I need to use knowledge of advanced EXCEL to draw meaningful conclusions about the IMDB Movies list.

Data Analytics Tasks:

A. **Movie Genre Analysis:** Analyze the distribution of movie genres and their impact on the IMDB score.
B. **Movie Duration Analysis:** Analyze the distribution of movie durations and its impact on the IMDB score.
C. **Language Analysis:** Situation: Examine the distribution of movies based on their language.
D. **Director Analysis:** Influence of directors on movie ratings.
E. **Budget Analysis:** Explore the relationship between movie budgets and their financial success.

My Approach

Firstly, I have taken the dataset and studied the whole dataset to gets some knowledge and get some idea to make decisions. After that, cleaned the dataset removed unnecessary things.

After analysis, I have started reading tasks understand the statement and identified required columns to solve the tasks. Started creating tables, pivot tables based on task and create graphs or charts to represent visualization of data. This is my approach to solve the given tasks.

Software used to do this project

Microsoft Office Excel

Insights

By doing this project, I have learned more about analysing the large dataset.

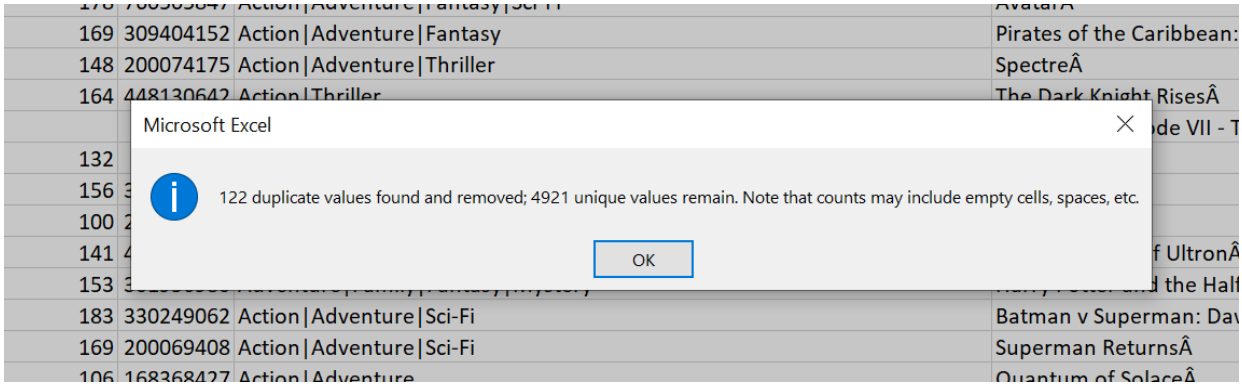
Data Cleaning

Cleaning data is the most important thing to perform analysis.

1.Removing unnecessary columns which are not required for analysis:

Color, num_critic_for_reviews, director_facebook_likes, actor_3_facebook_likes, actor_2_name, actor_1_facebook_likes, actor_1_name, num_voted_users, cast_total_facebook_likes, actor_3_name, facenumber_in_poster, plot_keywords, movie_imdb_link, num_user_for_reviews, content_rating, actor_2_facebook_likes, aspect_ratio and movie_facebook_likes.

2.Removing duplicate rows:



3.Removed rows which having blanks in column deleted entire rows.

4.In movie_title, at the end of the text in every cell there is unnecessary character is there "Â ". I have removed this using formula

=LEFT(E2,LEN(E2)-2)

Before cleaning process: **4998 rows**

After cleaning process: **3786 rows**

My Results

Data Analytics Tasks:

A. Movie Genre Analysis: To analyze the distribution of movie genres and their impact on the IMDB score.

To analyzing the distribution of movie genres and their impact on IMDB scores, used genres, movie title and imdb score columns.

Taken genres column to get unique genres list.

Genres
Action
Adventure
Drama
Animation
Comedy
Mystery
Crime
Biography
Fantasy
Documentary
Sci-Fi
Horror
Romance
Family
Western
Musical
Thriller
History
Music
War
Sport
Short
Film-Noir

Extracted genres, imdb_score and movie_title columns to do further process.

And taken the unique genres list as columns to create table for each movie, if genre is present among 23 genres, then it will get the genres name.

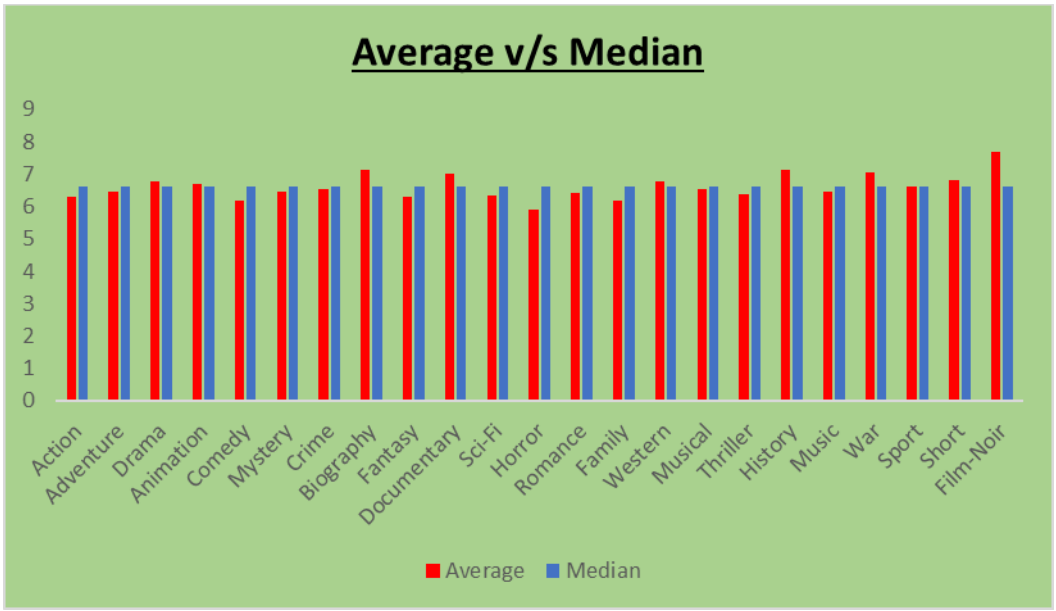
Using countifs function, finds the count of each genre.

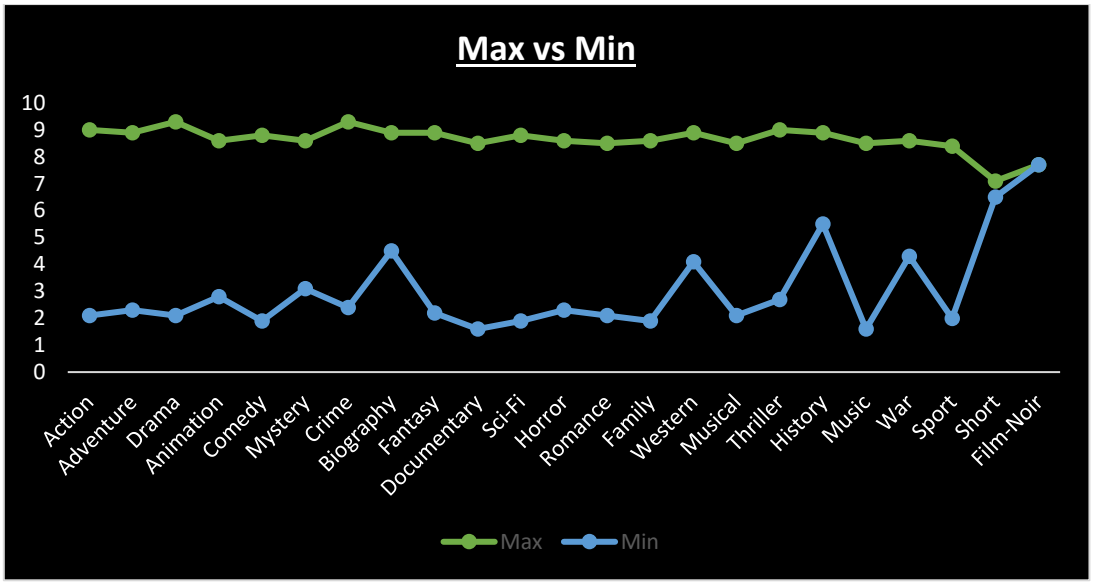
Using above created table, created pivot area and find average, median, maximum, minimum, variance and standard deviation of each genre.

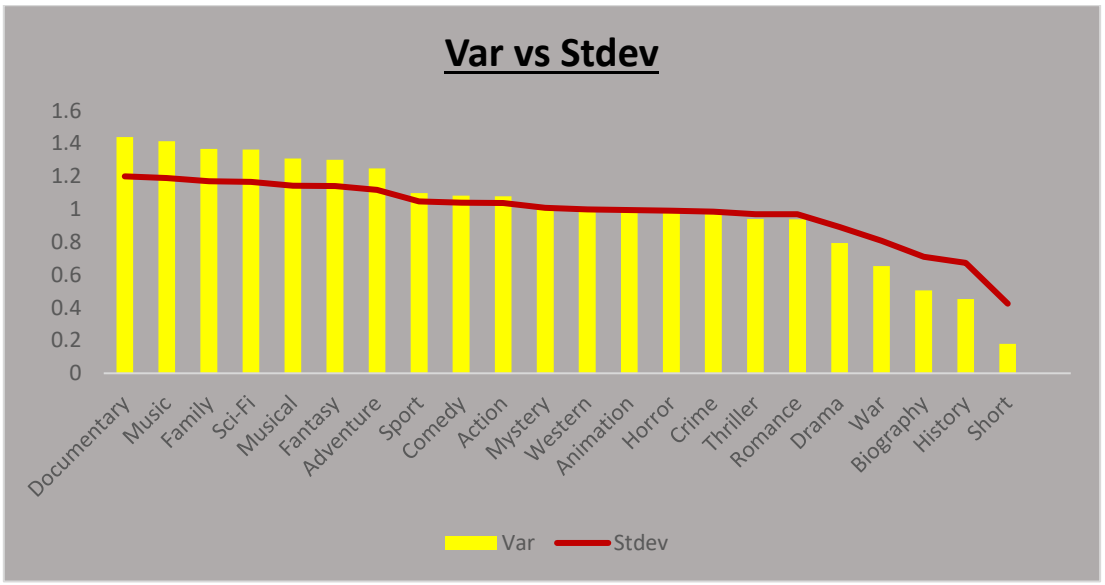
Table:

Genres	Average	Median	Max	Min	Var	Stdev
Action	6.285989305	6.6	9	2.1	1.078186788	1.038357736
Adventure	6.454960836	6.6	8.9	2.3	1.247524378	1.116926308
Drama	6.789115646	6.6	9.3	2.1	0.793996652	0.891064898
Animation	6.700507614	6.6	8.6	2.8	0.987295659	0.993627525
Comedy	6.183310992	6.6	8.8	1.9	1.081431552	1.039919012
Mystery	6.469496021	6.6	8.6	3.1	1.014838309	1.007391835
Crime	6.548148148	6.6	9.3	2.4	0.968463042	0.984105199
Biography	7.140082645	6.6	8.9	4.5	0.504237338	0.71009671
Fantasy	6.285080645	6.6	8.9	2.2	1.30054464	1.140414241
Documentary	7.011940299	6.6	8.5	1.6	1.439855269	1.199939694
Sci-Fi	6.327272727	6.6	8.8	1.9	1.362318841	1.16718415
Horror	5.903957784	6.6	8.6	2.3	0.982127152	0.991023285
Romance	6.426212471	6.6	8.5	2.1	0.938953731	0.968996249
Family	6.2	6.6	8.6	1.9	1.367909091	1.169576458
Western	6.765517241	6.6	8.9	4.1	0.997035693	0.998516746
Musical	6.550980392	6.6	8.5	2.1	1.307672297	1.143535
Thriller	6.372309108	6.6	9	2.7	0.939112803	0.969078327
History	7.131578947	6.6	8.9	5.5	0.451578947	0.671996241
Music	6.456680162	6.6	8.5	1.6	1.413359666	1.188848041
War	7.048427673	6.6	8.6	4.3	0.652386753	0.80770462
Sport	6.601360544	6.6	8.4	2	1.09876526	1.048220043
Short	6.8	6.6	7.1	6.5	0.18	0.424264069
Film-Noir	7.7	6.6	7.7	7.7	#NUM!	#NUM!

Chart:







B. Movie Duration Analysis: To analyze the distribution of movie durations and its impact on the IMDB score.

To find the average, median and standard deviation for movie durations, used excel functions of AVERAGE(), MEDIAN() and STDEV().

Formula:

=AVERAGE(A2:A3787)

=MEDIAN(A2:A3787)

=STDEV(A2:A3787)

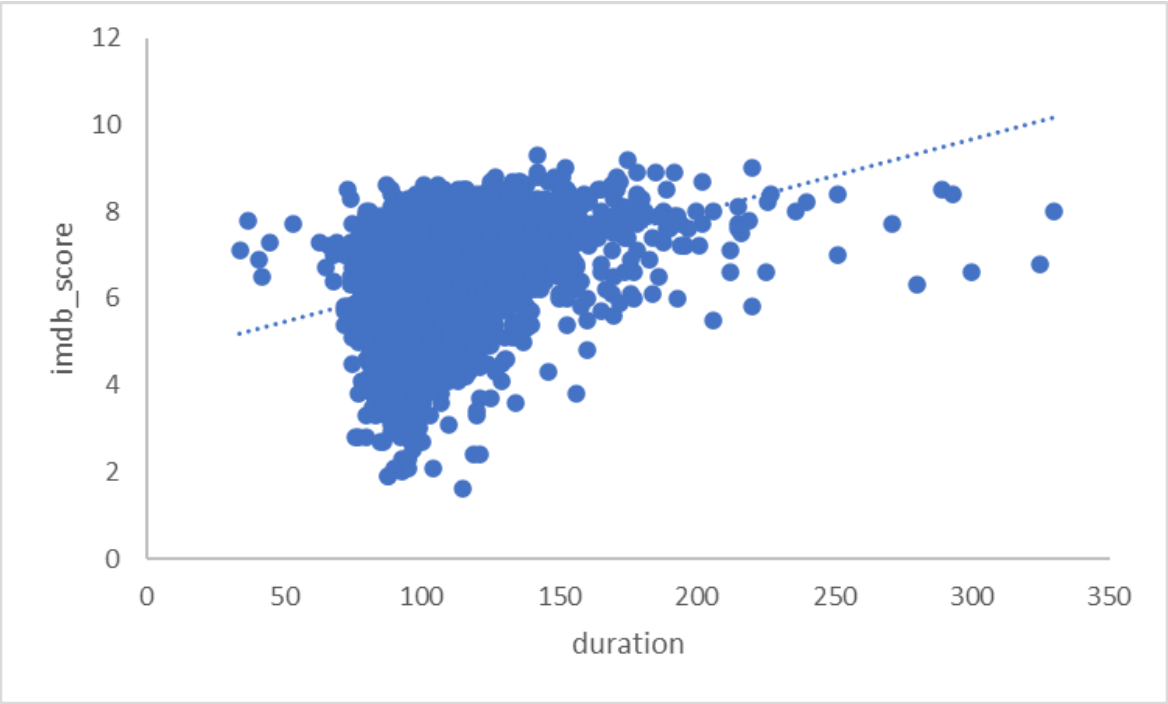
Output:

MEAN	109.8085
MEDIAN	105
STANDARD DEVIATION	22.7632

Blood In, Blood Out	330
Marilyn Hotchkiss' Ballroom Dancing and Charm School	34

Visualizing the relationship between movie duration and IMDB score.

Chart:



Added linear trendline it was moving upward.

C. Language Analysis: To show the distribution of movies based on their language.

To show the most common languages used in movies and analyzing their impact on the IMDB score using descriptive statistics.

First inserted pivot table with language as rows and calculated number of movies, mean, median and standard deviation of IMDB scores for each language by dragging the imdb_score column into the values and changed calculation type.

Table:

Language	Average of imdb_scor	Median of imdb_score	StdDevp of imdb_score
Aboriginal	6.95	6.95	0.55
Arabic	7.2	7.2	0
Aramaic	7.1	7.1	0
Bosnian	4.3	4.3	0
Cantonese	7.2375	7.3	0.412121038
Czech	7.4	7.4	0
Danish	7.9	8.1	0.43204938
Dari	7.5	7.5	0.1
Dutch	7.566666667	7.8	0.329983165
Dzongkha	7.5	7.5	0
English	6.421436495	6.5	1.052352956
Filipino	6.7	6.7	0
French	7.286486486	7.2	0.553691378
German	7.692307692	7.7	0.615769111
Hebrew	7.5	7.3	0.355902608
Hindi	6.76	7.05	1.05470375
Hungarian	7.1	7.1	0
Icelandic	6.9	6.9	0
Indonesian	7.9	7.9	0.3
Italian	7.185714286	7	1.069617517
Japanese	7.625	7.8	0.861321659
Kazakh	6	6	0
Korean	7.7	7.7	0.509901951
Mandarin	7.021428571	7.25	0.737930089
Maya	7.8	7.8	0
Mongolian	7.3	7.3	0
None	8.5	8.5	0
Norwegian	7.15	7.3	0.497493719
Persian	8.133333333	8.4	0.449691252
Portuguese	7.76	8	0.875442745
Romanian	7.9	7.9	0
Russian	6.5	6.5	0
Spanish	7.05	7.15	0.810151933
Swedish	7.6	7.6	0
Telugu	8.4	8.4	0
Thai	6.633333333	6.6	0.368178701
Vietnamese	7.4	7.4	0
Zulu	7.3	7.3	0

Chart:

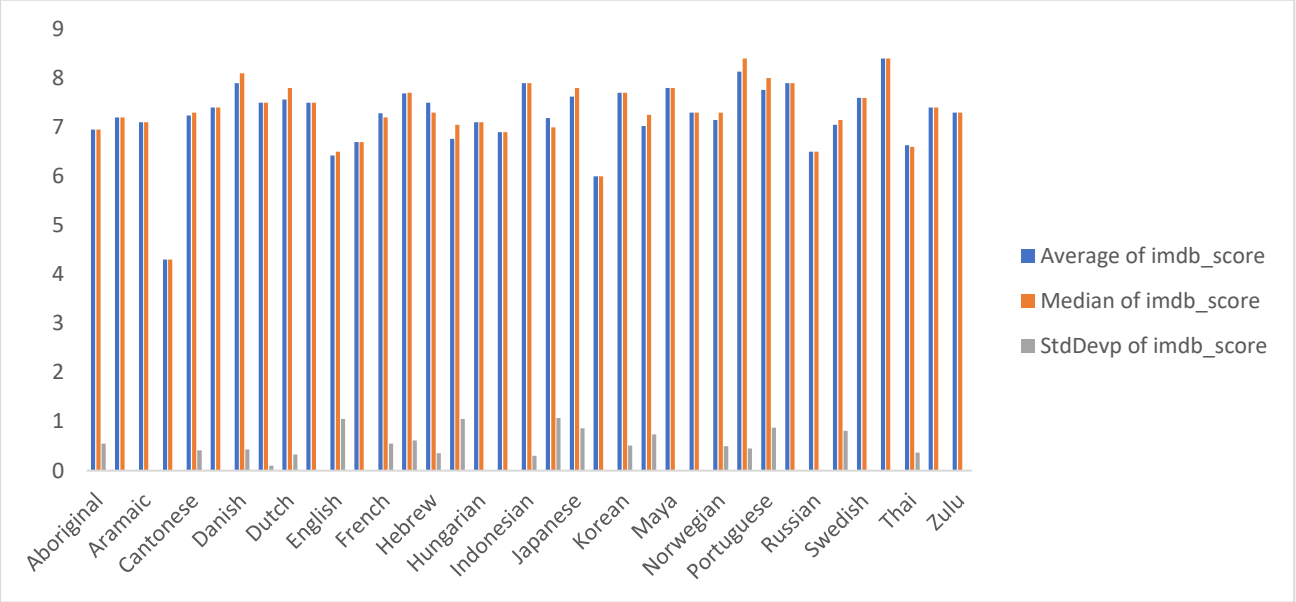
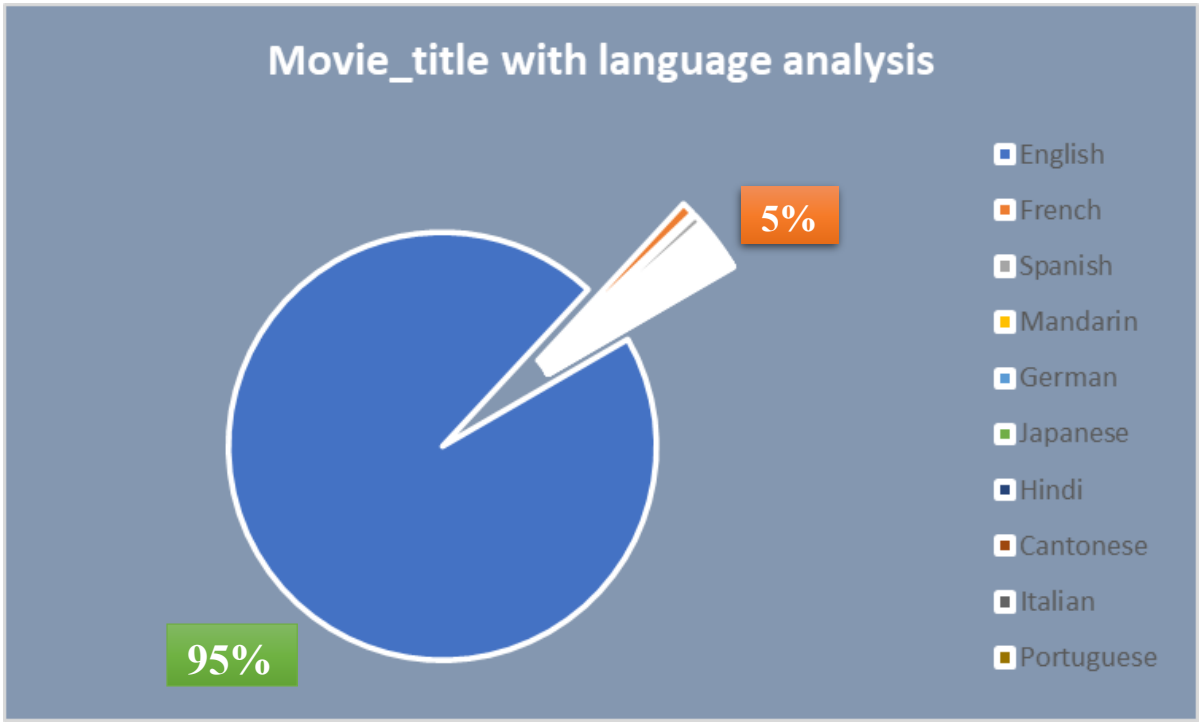


Table:

language	Count of movie_title
Aboriginal	2
Arabic	1
Aramaic	1
Bosnian	1
Cantonese	8
Czech	1
Danish	3
Dari	2
Dutch	3
Dzongkha	1
English	3606
Filipino	1
French	37
German	13
Hebrew	3
Hindi	10
Hungarian	1
Icelandic	1
Indonesian	2
Italian	7
Japanese	12
Kazakh	1
Korean	5
Mandarin	14
Maya	1
Mongolian	1
None	1
Norwegian	4
Persian	3
Portuguese	5
Romanian	1
Russian	1
Spanish	26
Swedish	1
Telugu	1
Thai	3
Vietnamese	1
Zulu	1

Chart:



For English, highest number movies with 95% percentage.

D. Director Analysis: To show the director with their movie ratings.

To find the top directors based on their average IMDB score with analyzing the contribution to the success of movies using percentile calculation.

First, created pivot table with dragging director_name into rows and imdb_score into values, changed the calculation type to average to get the average of imdb_score for each, same adding calculations of min and max columns and added count of imdb_score for each director using count function.

There are total 18 directors whole percentile is equal to or greater than 99%.

Table:

Director_name	Count of movie_title	Average of imdb_score	Min of imdb_score	Max of imdb_score	Percentile
Tony Kaye	1	8.6	8.6	8.6	99.9
Charles Chaplin	1	8.6	8.6	8.6	99.9
Ron Fricke	1	8.5	8.5	8.5	99.7
Majid Majidi	1	8.5	8.5	8.5	99.7
Damien Chazelle	1	8.5	8.5	8.5	99.7
Alfred Hitchcock	1	8.5	8.5	8.5	99.7
Sergio Leone	3	8.433333333	8	8.9	99.6
Christopher Nolan	8	8.425	7.2	9	99.6
S.S. Rajamouli	1	8.4	8.4	8.4	99.3
Richard Marquand	1	8.4	8.4	8.4	99.3
Marius A. Markevicius	1	8.4	8.4	8.4	99.3
Asghar Farhadi	1	8.4	8.4	8.4	99.3
Lee Unkrich	1	8.3	8.3	8.3	99.1
Lenny Abrahamson	1	8.3	8.3	8.3	99.1
Fritz Lang	1	8.3	8.3	8.3	99.1
Billy Wilder	1	8.3	8.3	8.3	99.1
Pete Docter	3	8.233333333	8.1	8.3	99
Hayao Miyazaki	4	8.225	7.7	8.6	99

E. Budget Analysis: To show the relationship between movie budgets and their financial success.

Taken the gross and budget columns into the separate sheet, calculated the profit formula (gross - budget) for each movie.

To calculate Correlation coefficient between gross and budget, used **=CORREL(A2:A3787,B2:B3787)** function.

Correlation coefficient	0.0965689

To find the highest profit earned movie used max() function for profit column.

=MAX(C2:C3787)

Maximum profit	523505847

By using this maximum value need to find the movie title, with using index and match functions.

=INDEX(D2:D3787,MATCH(G4,C2:C3787,0))

movie_title	Avatar

So, Avatar movie is the highest profit earned margin.

Drive link

IMDB_Movies.xlsx (Recommended)

IMDB Moive Excel sheet link

Conclusion

Here are the solutions given for the given tasks need to do as the data analyst.

In this task, used the concepts of statistics and advanced excel with basic and advanced topics to create tables and charts with statistics. These tables and charts were implemented using the Microsoft Office Excel.

THANK YOU.