

Name:- Yash Rajendra Gaikwad
Data Analytics Trainee
Project 5:- IMDB Movie Analysis.
Software Used:- Microsoft Excel.

❖ **Analysis done on following Points:-**

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 database. Then, for each genre, calculate descriptive statistics of the IMDB score.

B) Movie Duration Analysis:- Analyze the distribution of movie durations and its impact on IMDB score.

Task:- Analyse the distribution of movie durations and identify the relationship between duration and IMDB score.

C) Language Analysis:- Examine the distribution of movies based on their language.

Task:- Determine the most common language used in movies and analyze their impact on IMDB score.


D) Director Analysis:- Influence of directors on movie rating.

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

E) Budget Analysis:- Explore the relationship between movie budget and their financial success.

Task:- Analyze the correlation between movie budgets and gross earning and identify the movies with highest profit margin.

❖ Data Cleaning:-

- To clean the dataset we will be first dropping the columns which have no use for the analysis.
 - Columns like 'Color', 'director_facebook_likes', 'actor_3_facebook_likes', 'actor_2_name', 'actor_1_facebook_likes', 'cast_total_facebook_likes', 'actor_3_name', 'facenumber_in_posts', 'plot_keywords', 'movie_imdb_link', 'content_rating', 'actor_2_facebook_likes', 'aspect_ratio', 'movie_facebook_likes' are the columns containing irrelevant data for the analysis tasks provided. So, these columns need to be dropped.
 - After dropping the irrelevant columns now we need to remove the rows from the dataset having any of its column value as blank/NULL.
 - Then we need to get rid of the duplicate values in the dataset which can be achieved by using the 'Remove Duplicate Values/Cells' available in the 'Data' tab.
 - Now, all the data is cleaned and ready for analysis.
- 


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 database. Then, for each genre, calculate descriptive statistics of the IMDB score.

- **Process:-**

- Most of the movies contain more than one genre. so, use COUNTIF function to count the number of movies of each genre.
- After that, we use Pie Chart to show the movies of specific genre.
- For better clarification, we calculate the Median, Max, Min, Mode, Var and STDEV.
- Lastly compare it with the IMDB score.

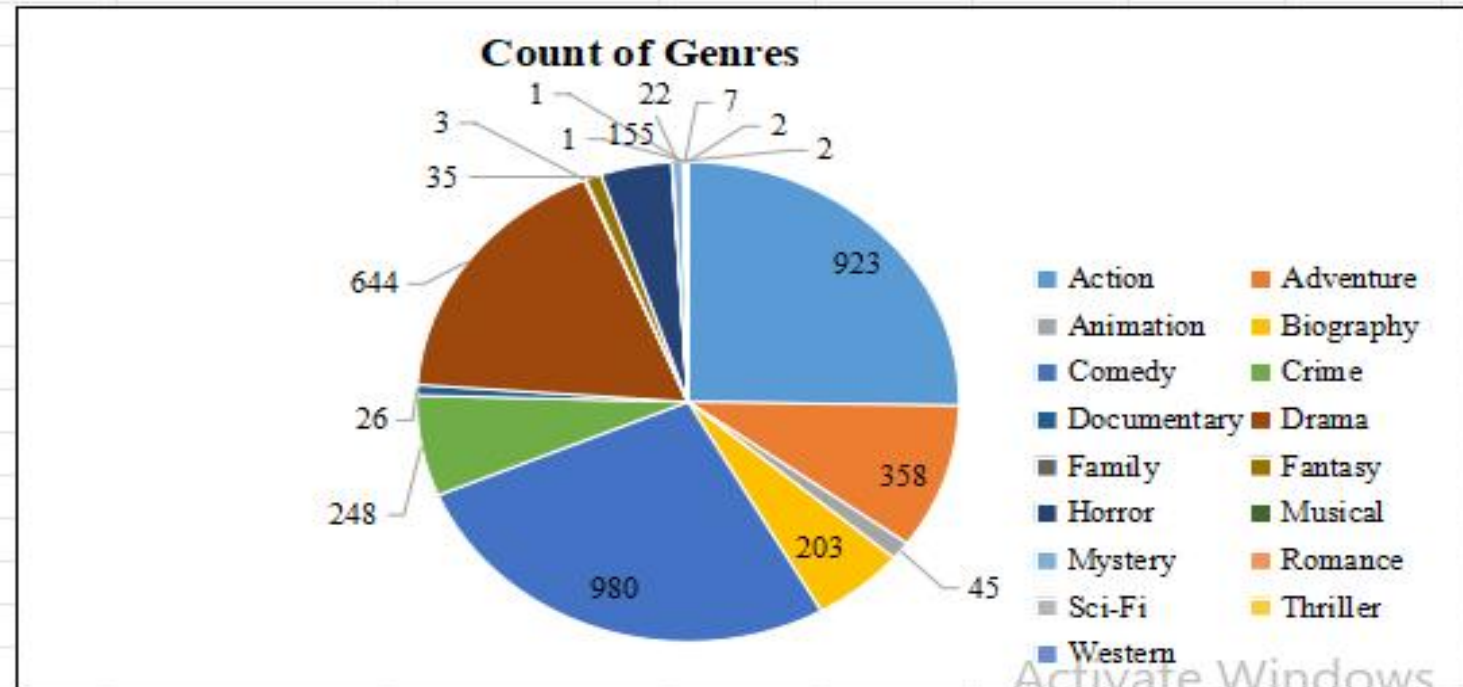
- **Result:-**

- As we can see in the below Action, Comedy, Drama, Adventure and crime are the most popular genres.
 - Top 20 movies according to IMDB score are mostly from Action, Drama, Crime Genres.
- 

❖ Result:-

Genres	Count of Genre
Action	923
Adventure	358
Animation	45
Biography	203
Comedy	980
Crime	248
Documentary	26
Drama	644
Family	3
Fantasy	35
Horror	155
Musical	2
Mystery	22
Romance	1
Sci-Fi	7
Thriller	1
Western	2

Genres Manipulation	
Median	30.5
Max	980
Min	1
Mode	2
Variance	73001.6875
Standard Deviation	270.1882446




B)Movie Duration Analysis:- Analyze the distribution of movie durations and its impact on IMDB score.

Task:- Analyse the distribution of movie durations and identify the relationship between duration and IMDB score.

- **Process:-**

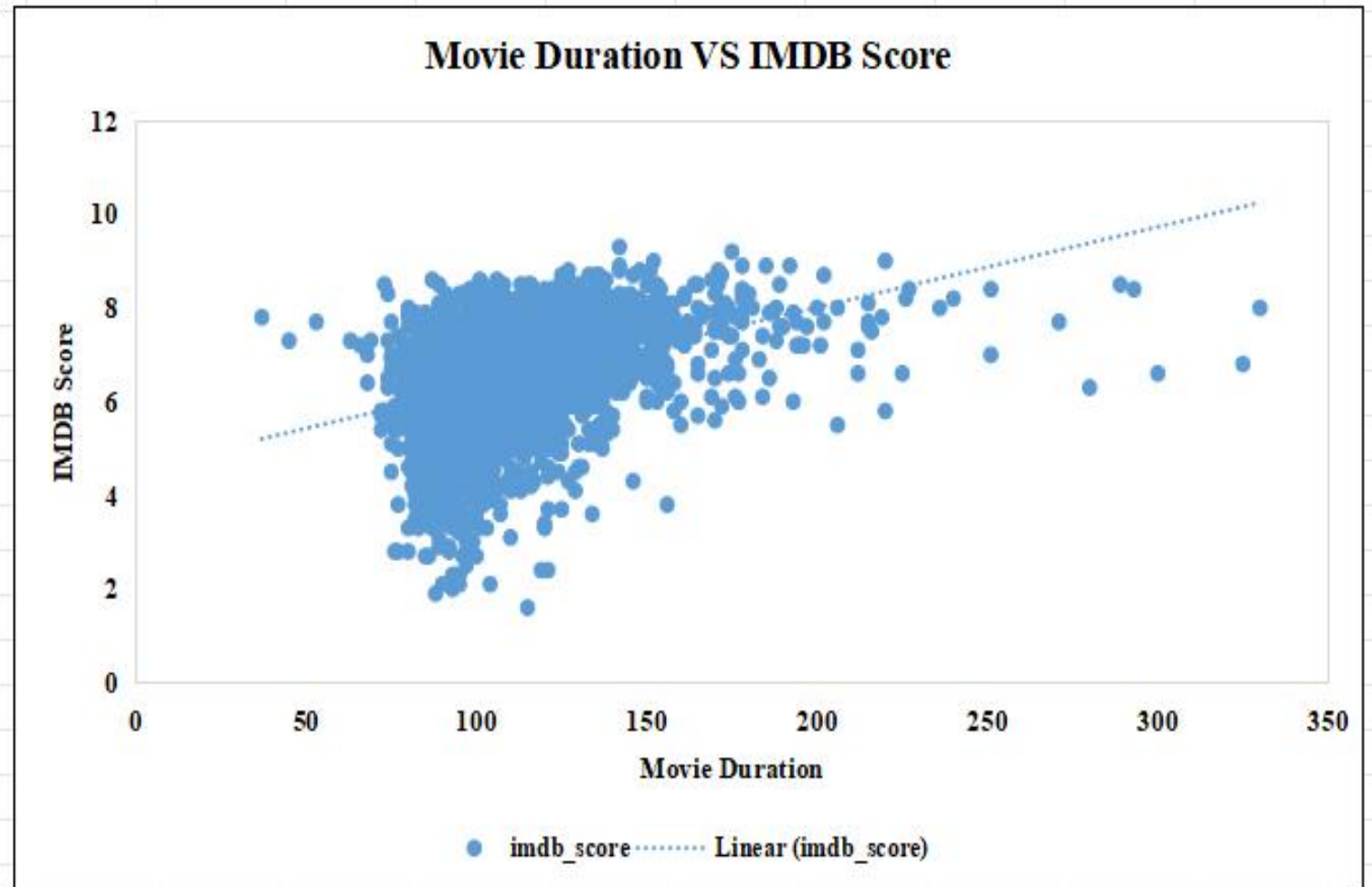
- First we take the duration column and the imdb_score Column in seprate sheet so we can understand it very clearly and our task should look clean.
- Then, we use the Scatter plot visualize the relationship between the movie duration and IMDB score.
- Also , we add the trendline to assess the direction and strength of the relationship.
- Lastly, we calculate the descriotive statistics such as Average, Median and Standard Deviation.

- **Result:-**

- In Scatter Plot we see that, most of the movies duration is between 75 - 175 min.
 - Most of the highest IMDB score movies duration is between 140 - 250 min.
 - Also most of the old movies have more movie duration then the new movies.
- 

❖ Result:-

Avg. Movie Duration	110.2
Median	106
Standard Deviation	22.7




C) Language Analysis:- Examine the distribution of movies based on their language.

Task:- Determine the most common language used in movies and analyze their impact on IMDB score.

- **Process:-**

- First we take the Movie Title, Language and IMDB score column on another sheet for better understanding.
- We use the Pivot Table option by placing the Language column in to the row and Count of Language into the values.
- We use the Pie Chart to visualize the movie language.
- After that, we use this Pivot Table and the IMDB score data to calculate the Mean, Median and the Standard Deviation for the all language.

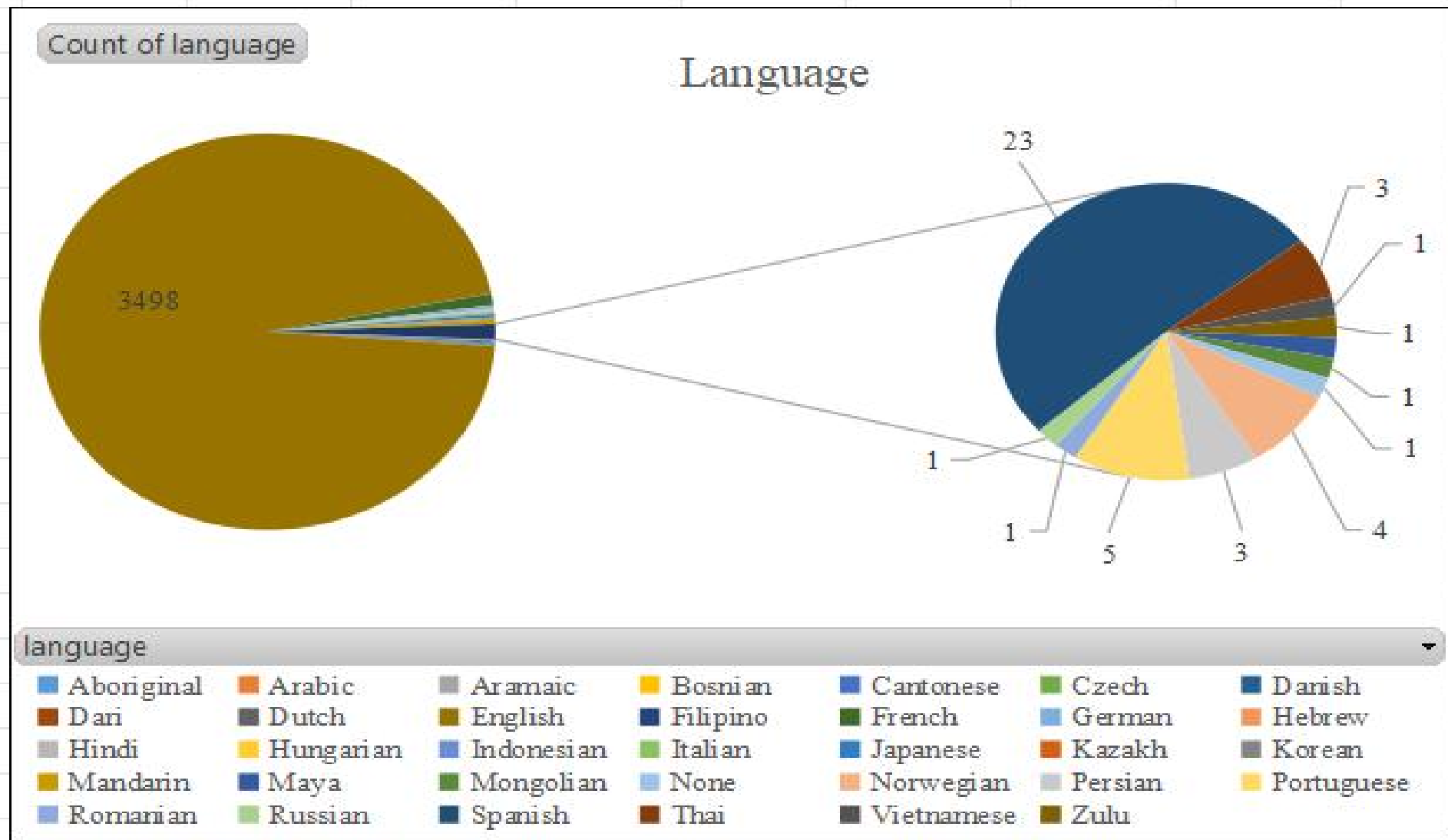
- **Result:-**

- Most of the movies are in English Language.
 - French and Spanish are the 2nd and 3rd languages used in movies.
 - All Top 20 IMDB score movies are English language movies.
- 

❖ Pivot Table and Mean, Median, STDEV Tables:-

language	Count of language	Language	Mean	Median	Standard Deviation
Aboriginal	2	Aboriginal	7.0	6.6	1.1
Arabic	1	Arabic	7.2	6.6	1.1
Aramaic	1	Aramaic	7.1	6.6	1.1
Bosnian	1	Bosnian	4.3	6.6	1.1
Cantonese	7	Cantonese	7.3	6.6	1.1
Czech	1	Czech	7.4	6.6	1.1
Danish	3	Danish	7.9	6.6	1.1
Dari	2	Dari	7.5	6.6	1.1
Dutch	3	Dutch	7.6	6.6	1.1
English	3498	English	6.4	6.6	1.1
Filipino	1	Filipino	6.7	6.6	1.1
French	34	French	7.4	6.6	1.1
German	10	German	7.8	6.6	1.1
Hebrew	1	Hebrew	8	6.6	1.1
Hindi	5	Hindi	7.2	6.6	1.1
Hungarian	1	Hungarian	7.1	6.6	1.1
Indonesian	2	Indonesian	7.9	6.6	1.1
Italian	7	Italian	7.2	6.6	1.1
Japanese	10	Japanese	7.7	6.6	1.1
Kazakh	1	Kazakh	6	6.6	1.1
Korean	5	Korean	7.7	6.6	1.1
Mandarin	14	Mandarin	7	6.6	1.1
Maya	1	Maya	7.8	6.6	1.1
Mongolian	1	Mongolian	7.3	6.6	1.1
None	1	None	8.5	6.6	1.1
Norwegian	4	Norwegian	7.2	6.6	1.1
Persian	3	Persian	8.1	6.6	1.1
Portuguese	5	Portuguese	7.8	6.6	1.1
Romanian	1	Romanian	7.9	6.6	1.1
Russian	1	Russian	6.5	6.6	1.1
Spanish	23	Spanish	7.1	6.6	1.1
Thai	3	Thai	6.6	6.6	1.1
Vietnamese	1	Vietnamese	7.4	6.6	1.1
Zulu	1	Zulu	7.3	6.6	1.1
Grand Total	3655				

❖ Result:-




D) Director Analysis:- Influence of directors on movie rating.

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

- **Process:-**

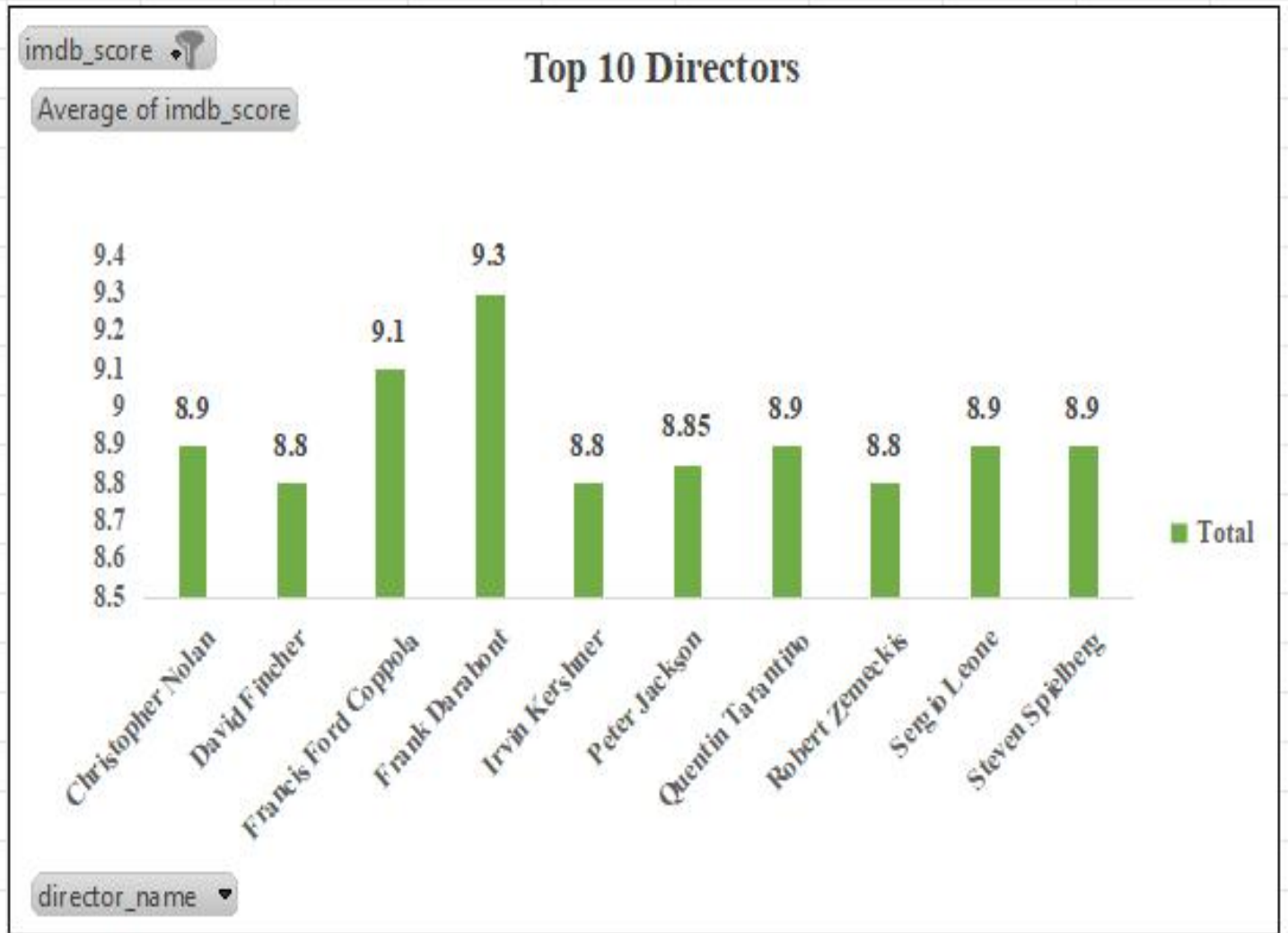
- It is very simple, first take the Director name and IMDB score column and use the Pivot Table option.
- We put the Director name in Row, IMDB score in filter and again IMDB score in values we use the Avg. of IMDB score option.
- We directly get the table. but we only use Top 10 Director for visualization.

- **Result:-**

- Frank Darabont have highest IMDB score followed by Francis Ford Coppola, Quentin tarantino, Christopher Nolan, Peter Jackson and David Fincher.
 - Frank Darabont have directed the highest IMDB scored movie i.e. The Shawshank Redemption.
 - Francis Ford Coppola's The Godfather movie series have 9 and 9.2 IMDB score.
- 

❖ Result:-

imdb_score	(Multiple Items)	T=
director_name	Average of imdb_score	
Christopher Nolan	8.9	
David Fincher	8.8	
Francis Ford Coppola	9.1	
Frank Darabont	9.3	
Irvin Kershner	8.8	
Peter Jackson	8.85	
Quentin Tarantino	8.9	
Robert Zemeckis	8.8	
Sergio Leone	8.9	
Steven Spielberg	8.9	
Grand Total	8.930769231	




E) Budget Analysis:- Explore the relationship between movie budget and their financial success.

Task:- Analyze the correlation between movie budgets and gross earnings and identify the movies with the highest profit margin.

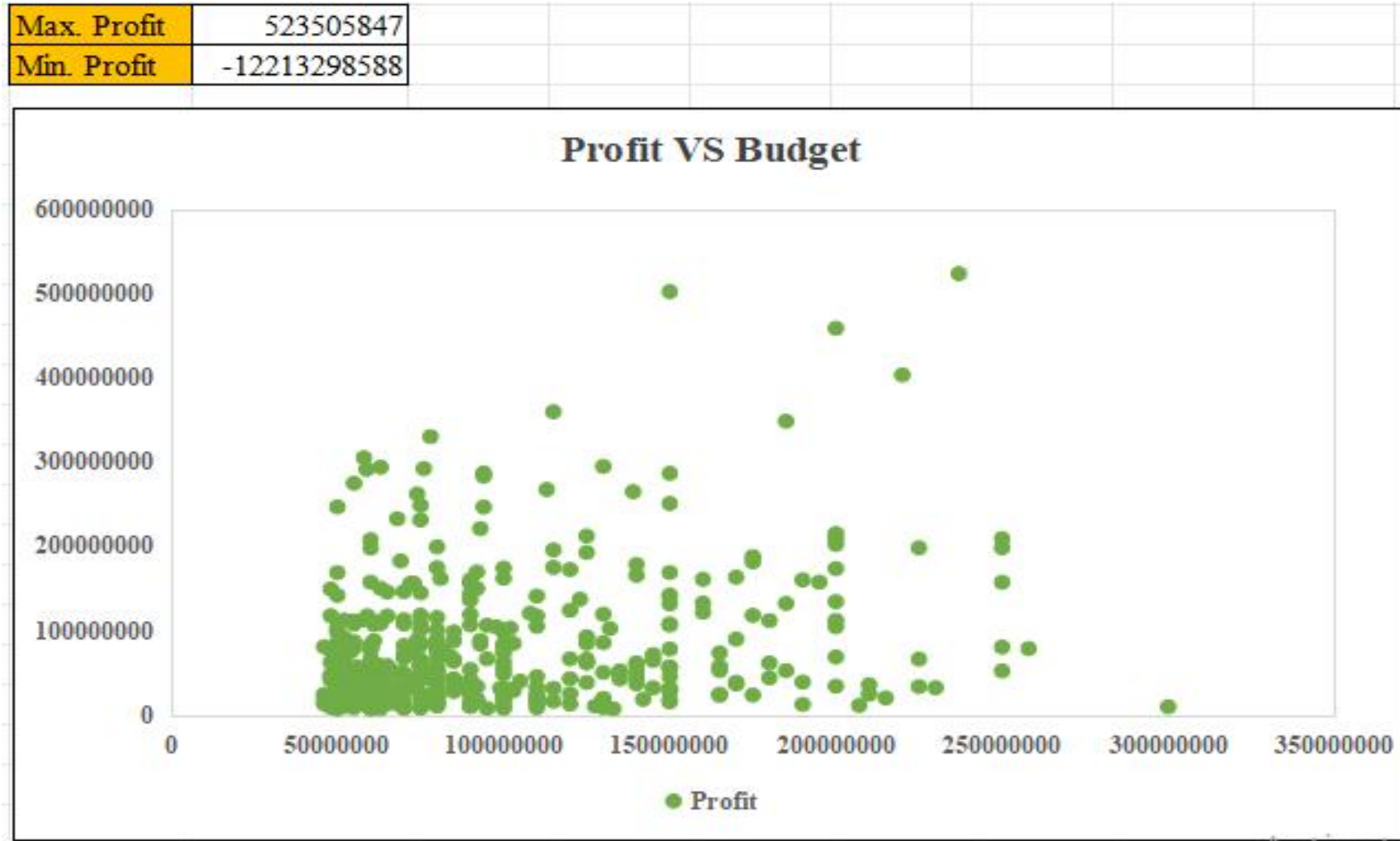
- **Process:-**

- First we select the Budget and Gross earnings column on another sheet to make it more simple.
- We minus the Budget from the Gross earnings to get the Profit for each movie.
- After that, we use this Budget column and Profit column for visualization.
- It will be more suitable to use the Scatter plot for this task.
- Also we calculate the Maximum and Minimum Profit by using the Max and Min function.

- **Result:-**

- Avatar is the most Profitable movie of all time followed by Jurassic World and Titanic.
 - Paranormal Activity is the most Profitable low budget movie of all time.
 - The Host is the least profitable movie.
- 

❖ Result:-



❖ **Additional Insight 1 :-** Top 20 IMDB Scored Movies.

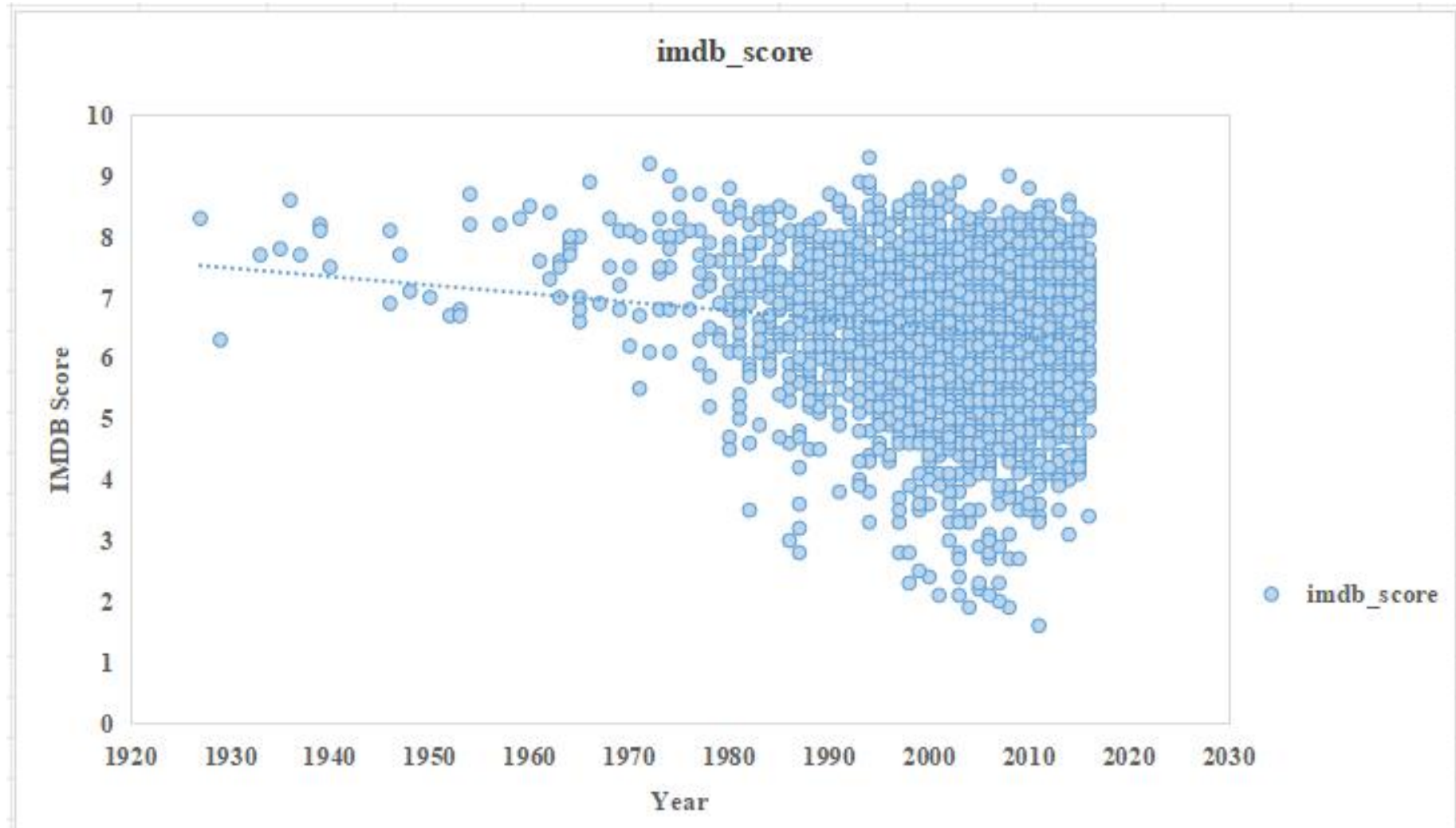
Result:- 1) The Shawshank Redemption movie have highest IMDB score.

2) The Lord of the Rings and Godfather is the highest IMDB Score Movie Series.

imdb_score	(Multiple Items)	T=
movie title	Average of imdb_score	
City of God		8.7
Fight Club		8.8
Forrest Gump		8.8
Goodfellas		8.7
Inception		8.8
One Flew Over the Cuckoo's Nest		8.7
Pulp Fiction		8.9
Schindler's List		8.9
Seven Samurai		8.7
Star Wars: Episode IV - A New Hope		8.7
Star Wars: Episode V - The Empire Strikes Back		8.8
The Dark Knight		9
The Godfather		9.2
The Godfather: Part II		9
The Good, the Bad and the Ugly		8.9
The Lord of the Rings: The Fellowship of the Ring		8.8
The Lord of the Rings: The Return of the King		8.9
The Lord of the Rings: The Two Towers		8.7
The Matrix		8.7
The Shawshank Redemption		9.3
Grand Total		8.85

❖ **Additional Insight 2 :-** IMDB Score VS Movie Release Year

- Result :-** 1) Most of the Old movies have high IMDB rating.
2) IMDB rating is declining as years passes.



END