IMDB Movie Analysis

Description –

This is data analytic project of IMDB movies. I completed the following task as leading team provided me by using my data analytics skills in excel. The tasks which I needed to complete are following –

1. Clean the data and make it more understandable and readable for better analysing.
2. The movies with highest profits.
3. The top 250 IMDB movies according to user ratings
4. The most popular genre of movies
5. The actors who is more popular in both critics audiences.

Approach -

I started my homework my learning more about excel analysis and how can I perform the necessary function according to the need of analysis. Then I started understanding the given dataset and tried to find the pattern or anything which can help me in cleaning the data. The I started my actual work my cleaning the dataset by performing certain steps and then started my analysis for given tasks.

Tech-stack used –

I used the window version of Microsoft excel.

Insights-

I used many excel functions and got to know more about excel features. I got more Insights on the project, like how can I perform same analysis with different tools and features in the excel and how it can be more understandable by doing different analysis.

Tasks –

A. **Cleaning the data -** This is one of the most important step to perform before moving forward with the analysis.

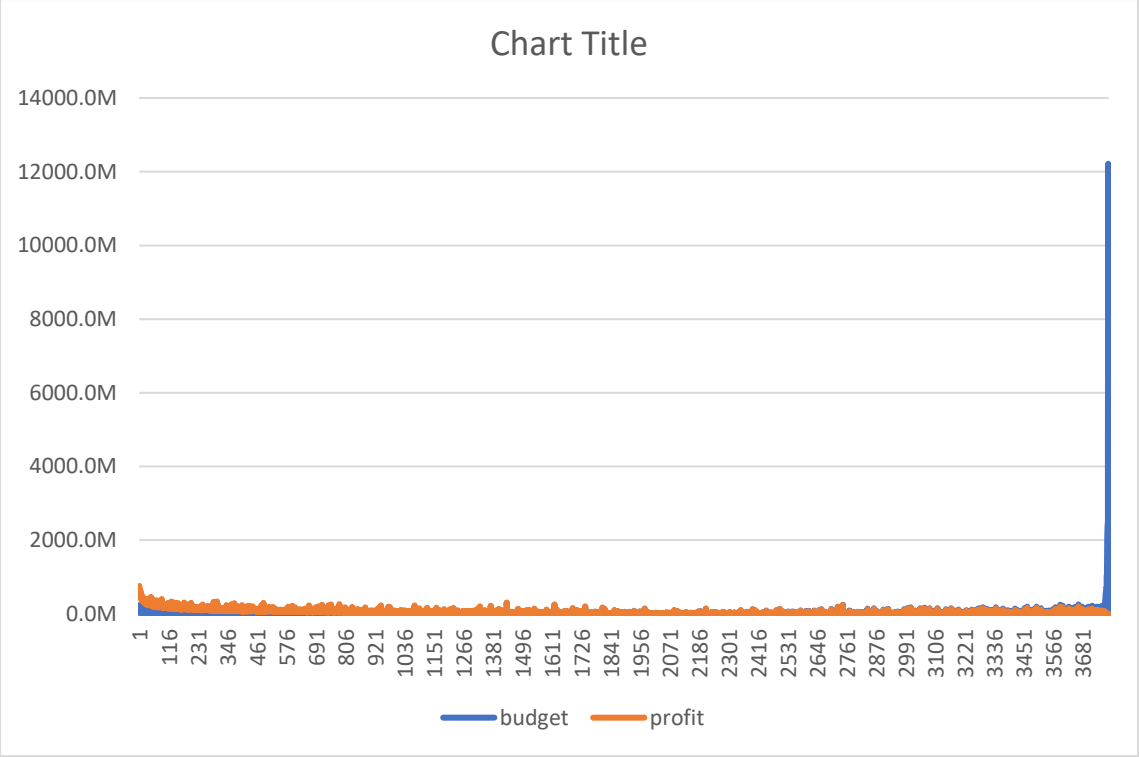
My task – Clean the data

1. I started my data cleaning by firstly understanding the pattern
2. I removed all the rows which contain empty or null columns.
3. I removed all the duplicate rows.
4. I removed the column which was not necessary during the main analysis.
5. I changed the column size and format, to make it more understandable and readable.
6. I arranged the column in the way which make it more readable and finding connection between columns.

B. **Movies with highest profit:** Create a new column called profit which contains the difference of the two columns: gross and budget. Sort the column using the profit column as reference. Plot profit (y-axis) vs budget (x-axis) and observe the outliers using the appropriate chart type.

My task: Find the movies with the highest profit?

1. Firstly, copy all the necessary column which are going to help in in task to the different worksheet for make it easy to approach the dataset.
2. Create a new column ‘profit’ which contain the difference of two column ‘budget’ and ‘gross’. Use the right function for finding the subtraction of two given column.
3. Sort the data in descending order as ‘profit’ column reference.
4. There you go, the required data for most profitable movies of all time.



	movie_title	gross	budget	profit
2	Avatar	760.5M	237.0M	523.5M
3	Jurassic World	652.2M	150.0M	502.2M
4	Titanic	658.7M	200.0M	458.7M
5	Star Wars: Episode IV - A New Hope	460.9M	11.0M	449.9M
6	E.T. the Extra-Terrestrial	434.9M	10.5M	424.4M
7	The Avengers	623.3M	220.0M	403.3M
8	The Lion King	422.8M	45.0M	377.8M
9	Star Wars: Episode I - The Phantom Menace	474.5M	115.0M	359.5M
10	The Dark Knight	533.3M	185.0M	348.3M
11	The Hunger Games	408.0M	78.0M	330.0M
12	Deadpool	363.0M	58.0M	305.0M
13	The Hunger Games: Catching Fire	424.6M	130.0M	294.6M
14	Jurassic Park	356.8M	63.0M	293.8M
15	Despicable Me 2	368.0M	76.0M	292.0M
16	American Sniper	350.1M	58.8M	291.3M
17	Finding Nemo	380.8M	94.0M	286.8M
18	Shrek 2	436.5M	150.0M	286.5M
19	The Lord of the Rings: The Return of the King	377.0M	94.0M	283.0M
20	Star Wars: Episode VI - Return of the Jedi	309.1M	32.5M	276.6M
21	Forrest Gump	329.7M	55.0M	274.7M
22	Star Wars: Episode V - The Empire Strikes Back	290.2M	18.0M	272.2M
23	Home Alone	285.8M	18.0M	267.8M
24	Star Wars: Episode III - Revenge of the Sith	380.3M	113.0M	267.3M
25	Spider-Man	403.7M	139.0M	264.7M
26	Minions	336.0M	74.0M	262.0M
27	The Sixth Sense	293.5M	40.0M	253.5M
28	Jaws	260.0M	8.0M	252.0M
29	Frozen	400.7M	150.0M	250.7M
30	The Secret Life of Pets	323.5M	75.0M	248.5M

C. **Top 250:** Create a new column IMDb_Top_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb_score). Also make sure that for all of these movies, the num_voted_users is greater than 25,000. Also add a Rank column containing the values 1 to 250 indicating the ranks of the corresponding films.

Extract all the movies in the IMDb_Top_250 column which are not in the English language and store them in a new column
named Top_Foreign_Lang_Film. You can use your own imagination also

My task: Find IMDB Top 250

1. Firstly, copy all the necessary column which is going to help in the analysis of this task.
2. Add the new column ‘rank’ before ‘movie title’ column and use the ROW()-1 function for ranking all the movies tittle.
3. Filter the imdb_score in descending order and also, filter the num_voted_user on number filter basis with value greater then 25000.
4. Movies title will be sorted automatically in a required order then copy all column to another sheet till 250 rows only.
5. Now, we got the required data for this task.
6. Now, the next step is to find all the popular movies which is not in ‘english’ language and that can be done with filtering the imdb top 250 movies.
7. The result of this filtering can be copy into new column ‘top foreign movies’

Rank	movie_title	language	imdb_score
1	The Shawshank Redemption	English	9.3
2	The Godfather	English	9.2
3	The Dark Knight	English	9
4	The Godfather: Part II	English	9
5	The Lord of the Rings: The Return of the King	English	8.9
6	Pulp Fiction	English	8.9
7	Schindler's List	English	8.9
8	The Good, the Bad and the Ugly	Italian	8.9
9	Forrest Gump	English	8.8
10	Star Wars: Episode V - The Empire Strikes Back	English	8.8
11	The Lord of the Rings: The Fellowship of the Ring	English	8.8
12	Inception	English	8.8
13	Fight Club	English	8.8
14	Star Wars: Episode IV - A New Hope	English	8.7
15	The Lord of the Rings: The Two Towers	English	8.7
16	The Matrix	English	8.7
17	One Flew Over the Cuckoo's Nest	English	8.7
18	Goodfellas	English	8.7
19	City of God	Portuguese	8.7
20	Seven Samurai	Japanese	8.7
21	Saving Private Ryan	English	8.6
22	The Silence of the Lambs	English	8.6
23	Se7en	English	8.6
24	Interstellar	English	8.6
25	The Usual Suspects	English	8.6
26	American History X	English	8.6
27	Modern Times	English	8.6
28	Spirited Away	Japanese	8.6
29	The Lion King	English	8.5
30	Raiders of the Lost Ark	English	8.5
31	The Dark Knight Rises	English	8.5

Sheet1 | Sheet3 | Sheet2 | Sheet4 | **Top-250-IMDB Movies**

movie_title	language	imdb_score
The Good, the Bad and the Ugly	Italian	8.9
City of God	Portuguese	8.7
Seven Samurai	Japanese	8.7
Spirited Away	Japanese	8.6
The Lives of Others	German	8.5
Children of Heaven	Persian	8.5
Samsara	None	8.5
A Separation	Persian	8.4
Oldboy	Korean	8.4
Das Boot	German	8.4
Baahubali: The Beginning	Telugu	8.4
Amélie	French	8.4
Princess Mononoke	Japanese	8.4
The Hunt	Danish	8.3
Metropolis	German	8.3
Downfall	German	8.3
Pan's Labyrinth	Spanish	8.2
The Secret in Their Eyes	Spanish	8.2
Incendies	French	8.2
The Act of Killing	Indonesian	8.2
Howl's Moving Castle	Japanese	8.2
Amores Perros	Spanish	8.1
The Celebration	Danish	8.1
Elite Squad	Portuguese	8.1
The Sea Inside	Spanish	8.1
Tae Guk Gi: The Brotherhood of War	Korean	8.1
Akira	Japanese	8.1
A Fistful of Dollars	Italian	8
Central Station	Portuguese	8
Waltz with Bashir	Hebrew	8
Persepolis	French	8

Sheet1 | Sheet3 | Sheet2 | Sheet4 | **Top-250-foriegn IMDB_Movies**

D. **Best Directors:** Group the column using the director_name column.

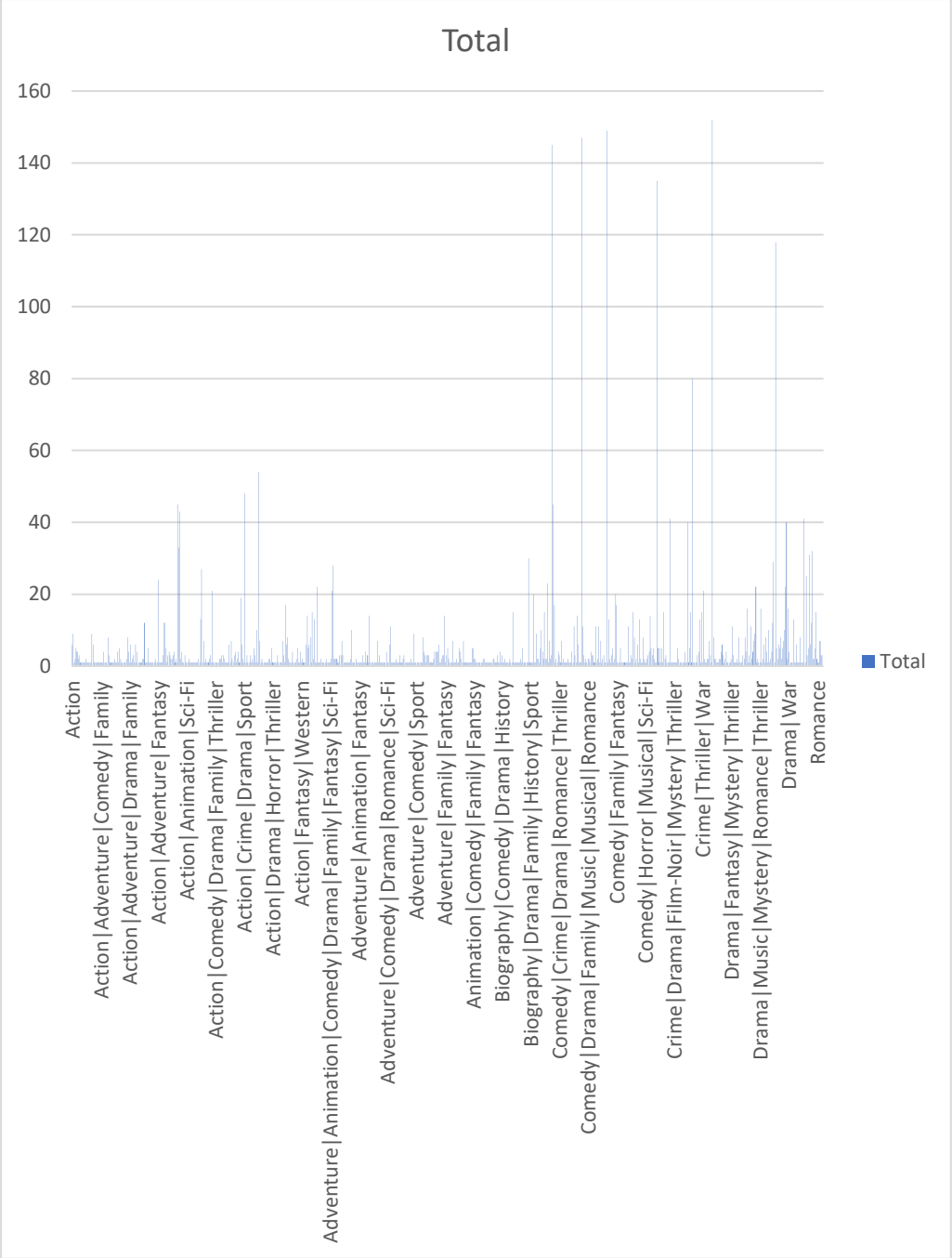
   Find out the top 10 directors for whom the mean of imdb_score is the highest and store them in a new column top10director. In case of a tie in IMDb score between two directors, sort them alphabetically.
   My task: Find the best directors

1. Start with making copy of all the required column in new sheet for analysis.
2. Make a new column ‘mean of imdb score’ and find the mean of same column and use the function AVERAGEIF for this purpose.
3. Next step is sorted this new column in descending order and filter the director column for any duplicate for this purpose and limit it for 10. And you got the top 10 director as asked by this task.

director_name	imdb_score	mean
Tony Kaye	8.6	8.6
Charles Chaplin	8.6	8.6
Ron Fricke	8.5	8.5
Damien Chazelle	8.5	8.5
Alfred Hitchcock	8.5	8.5
Majid Majidi	8.5	8.5
Sergio Leone	8	8
Christopher Nolan	8.5	8.2666667
Richard Marquand	8.4	8.4
S.S. Rajamouli	8.4	8.4

E. **Popular Genres:** Perform this step using the knowledge gained while performing previous steps.
My task: Find popular genre.

1. For this particular task, I used the pivot chart.
2. Pivot chart can easy analyse this column ‘ genre’ which contain very different content and pivot chart can easily do this analyse.
3. Select the column ‘genre’ and go to pivot chart function and make the suitable chart using this feature.



F. **Charts:** Create three new columns namely, Meryl_Streep, Leo_Caprio, and Brad_Pitt which contain the movies in which the actors: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors. Use only the actor_1_name column for extraction. Also, make sure that you use the names 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' for the said extraction.

Append the rows of all these columns and store them in a new column named Combined.

Group the combined column using the actor_1_name column.

Find the mean of
the num_critic_for_reviews and num_users_for_review and identify the actors which have the highest mean.

Observe the change in number of voted users over decades using a bar chart. Create a column called decade which represents the decade to which every movie belongs to. For example,
the title_year year 1923, 1925 should be stored as 1920s. Sort the column based on the column decade, group it by decade and find the sum of users voted in each decade. Store this in a new data frame called df_by_decade.

Your task: Find the critic-favorite and audience-favorite actors

1. Firstly, select all the necessary column related to this tasks to another sheet for analyse. Sort the data for all the three actor and put accordingly for all three actors.
2. Use the pivot chart for analysing the data and find the favorite actor both critic and audience. Change the data to ‘average’ in the value block of new window when we make pivot table.
3. So, by doing the analysing with chart, we got the answer to the given question perfectly.

	meryl streep	leonardo dicaprio	brad pitt
num_user_for_reviews	297.1818182	922.55	742.353
num_critic_for_reviews	181.4545455	322.2	245