IMDB Movie Analysis

Muskan Bansal

**Project Description:**

In this project I have a dataset having all the information I would need to perform the analysis and answer the questions, columns names are as follows:

Color: Tells if the movie is in Black & white film or color.

Director\_name: Has names of all the directors.

Num\_critic\_for\_reviews: Tells the number of critics.

Duration: How long is the movie.

Director\_facebook\_likes: Number of likes a director has on facebook.

Actro\_3\_facebook\_likes: Number of likes an actor has on facebook.

Actor\_2\_name: Name of the second actor.

Actor\_1\_facebook\_likes: Number of likes an actor has on facebook.

Gross: Gross is the sum of budget and profit.

Genres: Genere of the movie.

Actor\_1\_name: Name of the actor.

Movie\_title: Nmme of the movie.

Num\_voted\_users: Number of people voted.

Cast\_total\_facebook\_likes: Total likes on facebook of the cast members.

Actor\_3\_name: Nmae of actor.

Facenumber\_in\_poster: Number of faces on the movie poster.

Plot\_keywords: Key words to help describe the movie better.

Movie\_imdb\_link: IMDB Link of the movies.

Num\_user\_for\_reviews: Number of users.

Language: Language of the movie.

Country: Has name of the countries.

Content\_rating: Rating of the movie.

Budget: How much money is spend on the movie.

Title\_year: Year on which it was released.

Actor\_2\_facbook\_likes: Number of likes on facebook.

Imdb\_score: Score on IMDB.

Aspect\_ratio: The ratio of the width to the height of an image or screen.

Movie\_facebook\_likes: Number of likes a movie got on facebook.

**Approach:**

1. Understand the dataset provieded, load the data and import on excel sheet. Clean the data, remove the null values where needed, remove the outliers, remove the duplicate values where needed, convert the columns in the correct datatype, and create a clean table.

2. Understand the questions and what needs to be answered, make clean and need self-exploratory visualisations (Pi-Charts, Bar-diagrams, etc.) that helps to answer the questions easy.

3. Make a clean and need report. Using 5 Whys Analysis in the analysis and using this to create a report which conveys a data story. Five 'Whys' approach: It's also called the Root Cause Analysis, developed by Sakichi Toyoda, founder of Toyota Industries.

4. Deployment.

**Tech-Stack Used:**

**Excel and Word (Microsoft Office Home and Student 2019) for the purpose of visualisations and report making.**

**QUESTIONS:**

**QUESTION1:**

**Cleaning the data::** This is one of the most important step to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)  
**Your task:** Clean the data.

ANSWER: I cleaned the data, removing null values and duplicate data where necessary and created a clean table with correct datatypes.

**QUESTION2:**

**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.  
**Your task:** Find the movies with the highest profit?

ANSWER:

The name of the sheet in the excel is “Question2” where I copied the movie\_title, budget and gross columns and the I created another column “Profit” that stores the difference between budget and gross. The I selected the top 20 movies with the highest profit margin and plotted a bar graph for the same.

Then as was asked in the question, I also made a scatter chart that is also known as bubble chart where on the Y axis is the profit and that on the x axis is the budget., the scatter chart is a nice way to visualise and check for the outliers.

While The Avengers has the highest profit margin, Start wars and E.T the extra terrestrial are the two top movies with lowest budget and highest profit.

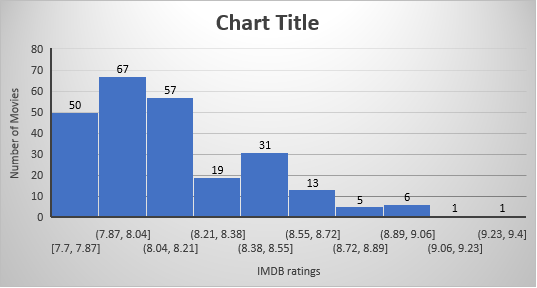
**QUESTION3:**

**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!  
**Your task:**Find IMDB Top 250.

ANSWERS:

Fisrt I copied the required columns on a new sheet and then I first ranked the num\_Voted\_users and removed all the datasets with value less than 25,000, then I also ranked the table on the basis of its IMDB ratings.

The above graph shows the top movies(including all the languages) with highest to low IMDB ratings and the length of the bars represents the number of voted users.

The above graph shows that most of the movies got ratings between 7.87-8.04.

The abover bar graph tells the highest to low IMDB ratings, 9.3 being the highest.(Including all the languages)

The abover bar graph tells the highest to low IMDB ratings, 9.3 being the highest.(Including only english movies.)

The abover bar graph tells the highest to low IMDB ratings, 9.3 being the highest.(Including all non-english movies.)

**QUESTION4:**

**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.

**Your task:**Find the best directors.

ANSWER:

First I copied the necessary columns in a new sheet and named it as Question4, I copied the columns movie\_title, director\_name and indb\_score.

Then I arranged the table on the basis of IMDB ratings in ascending order, calculated the mean IMDB rating which was 6.48 and removed all the data having ratings below that.

The above bar graph represents the names of all the directors with the top 10 ratings. Frank Darabot with the highest rating of 9.3.

**QUESTION5:**

**Popular Genres:**Perform this step using the knowledge gained while performing previous steps.  
**Your task:**Find popular genres.

ANSWER:

The above bar graph shows all the genres in an descending order where drama being the most popular genre out of the all.

The above is a pie representation of the generes where drama is the most popular with 22% and action at 4%.

The above bar graph represents the top 10 most popular genres.

**QUESTION6:**

**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.

ANSWER:

First, I copied all the necessary columns required for the visualisations and answering the questions in a new sheet and named it as question6.The I created a new table that only had names of the movies done by the 3 actors as mentioned.

The above pie chart explaining how many movies donw by the 3 characters.

Then I created a new sheet copying the necessary columns and named it as Q6 where I arranged names of the actors in the descending order on the basis of number of critics and number of users for the reviews and made the below 2 bar graph.

This one represents names of the actors in the descending order on the basis of number of users for reviews. And then the below bar graph represents the names of the cators in the descending order on the basis of number of critic for reviews.

Critic-favourite actor is Tom Hardy.

Audience-favourite actor is Christian Bale.

Finally again creating a new sheet and copying necessary columns, the I converted years into decade measure as was asked in the question and named the column as df\_by\_decade. I named this sheet as Q6.2.

And you can see that the number of users voting increased with time being 2020s as the most voted.