REPORT

Description – IMDb is a popular online database for movies and TV shows, providing comprehensive information about films, actors, directors, and more. It offers ratings, reviews, and trailers, serving as a go-to resource for film enthusiasts and industry professionals. Here we are provided with IMDB's dataset for movies from 1920-2010 which contains information about movies, actors, directors, budget, collection, ratings etc. We are going to clean the dataset and answer each question by using Five Why method for analytics using Microsoft Excel.

Approach – For this project, we will be understanding the data given to us. Then we will proceed to clean the data by getting rid of unnecessary columns, duplicates, blank values etc. After the cleaning, we'll use pivot tables, various functions and charts to answer certain questions.

Tech Stack used – Microsoft Excel

Insights –

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

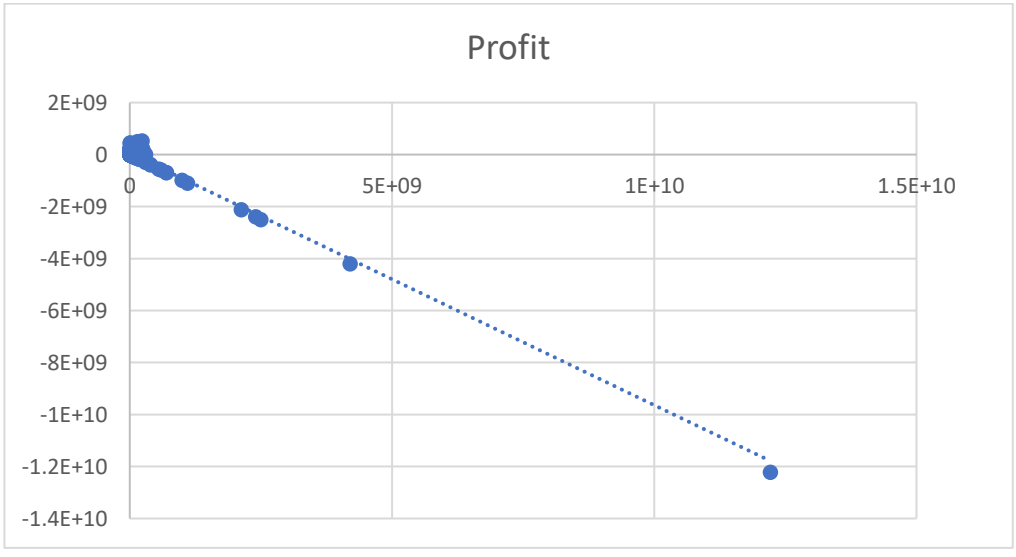
1. Dropping unnecessary columns – Color, director_facebook_likes, actor_3_facebook_likes, actor_2_name, actor_1_facebook_likes, cast_total_facebook_likes, actor_3_name, facenumber_in_posts, plot_keywords, movie_imdb_link, content_rating, actor_2_facebook_likes, aspect_ratio, movie_facebook_likes
2. Remove blank cell/null values
3. Remove duplicates

Cleaned Data: IMDB_Movies_Cleaned

B. Movies with highest profit: Create a new column called Profit which contains the difference of 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.

Task: Find the movies with the highest profit.

X-Y Scatter:



Outliers:

-12213298588

-4199788333

-2499804112

-2397701809

-2127109510

Top 5 Profitable Movies:

director_name	movie_title	budget	title_year	Profit
James Cameron	Avatar	237000000	2009	523505847
James Cameron	Titanic	200000000	1997	458672302
Colin Trevorrow	Jurassic World	150000000	2015	502177271
George Lucas	Star Wars: Episode IV - A New Hope	11000000	1977	449935665
Steven Spielberg	E.T. the Extra-Terrestrial	10500000	1982	424449459

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.

Task: Find IMDB Top 250

Top 250 English Movies: Top 250 IMDB English Movies

Top Foreign Movies: Top Foreign Movies

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

Task: Find the best directors

Director	Average IMDB Score
Alfred Hitchcock	8.5
Asghar Farhadi	8.4
Charles Chaplin	8.6
Christopher Nolan	8.4
Damien Chazelle	8.5
Majid Majidi	8.5
Marius A. Markevicius	8.4
Richard Marquand	8.4
Ron Fricke	8.5
S.S. Rajamouli	8.4

E. **Popular Genres:** Perform this step using the knowledge gained while performing previous steps.

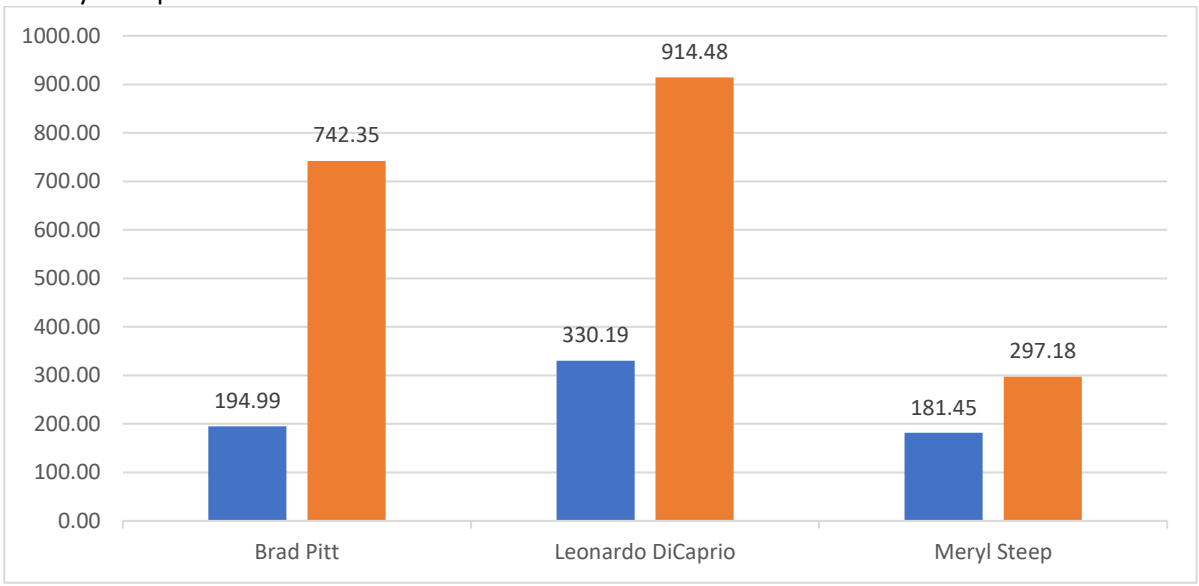
Your task: Find popular genres.

Genre	Count
Drama	148
Crime|Drama	147
Action|Adventure|Drama|Fantasy	147
Comedy|Drama	145
Crime|Drama|Thriller	82
Comedy|Romance	134
Drama|Romance	119
Comedy|Drama|Romance	151
Comedy	144
Action|Crime|Thriller	55

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.

Task: Find the critic-favorite and audience-favorite actors. Find the mean of the num_critic_for_reviews and num_users_for_review and identify the actors which have the highest mean.

actor_1_name	Mean of Critic Reviews	Mean of User Reviews
Brad Pitt	194.99	742.35
Leonardo DiCaprio	330.19	914.48
Meryl Steep	181.45	297.18



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.

Decade	Sum of num_voted_users
1920	116392
1930	804839
1940	230838
1950	678336
1960	2985581
1970	8704723
1980	20101705
1990	70090204
2000	173033966
2010	122492496