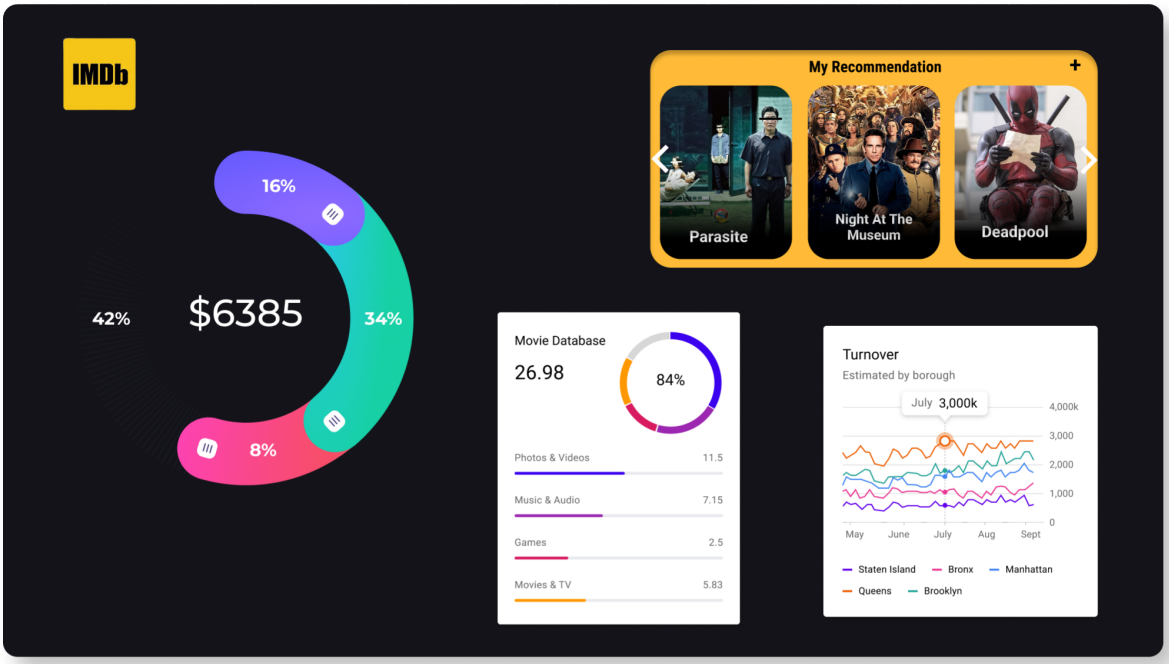
PROJECT-5
IMDB MOVIE ANALYSIS



PROJECT DESCRIPTION

We are providing you with a dataset having various columns of different IMDB Movies. You are required to Frame the problem. For this task, you will need to define a problem you want to shed some light on.

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.)

Once you have defined a problem, clean the data as necessary, and use your Data Analysis skills to explore the data set and derive insights.

Once you have framed the problem and gathered initial insights from the data, you can ask the following questions as you dig deeper into your analysis.

- What do you see happening?
- What are the specific symptoms of the problem?
- What is your hypothesis for the cause of the problem?

APPROACH

For completion of the project:

(1) Download all the data provided
(2) Understanding the data
(3) Finding duplicate & null values
(4) Data processing and solving the asked problems
(5) Create charts for easy representation

TECH-STACK USED



INSIGHTS

(A) Cleaning The Data

Inspect Null values using COUNTBLANK() function to count the number of empty cells in all the columns and rows.

Here, we are analysing the movies with respect to gross collection, ratings, popularity, etc and many columns are not required in the dataset. like : colour, director_facebook_likes, actor_1_facebook_likes, actor_2_facebook_likes, actor_3_facebook_likes, actor_2_name, cast_total_facebook_likes, actor_3_name, duration, facenumber_in_poster, content_rating, country, movie_imdb_link, aspect_ratio, plot_keywords.

We can also see that some columns have large percent of null values, which will drop such rows.

Removing Duplicate.

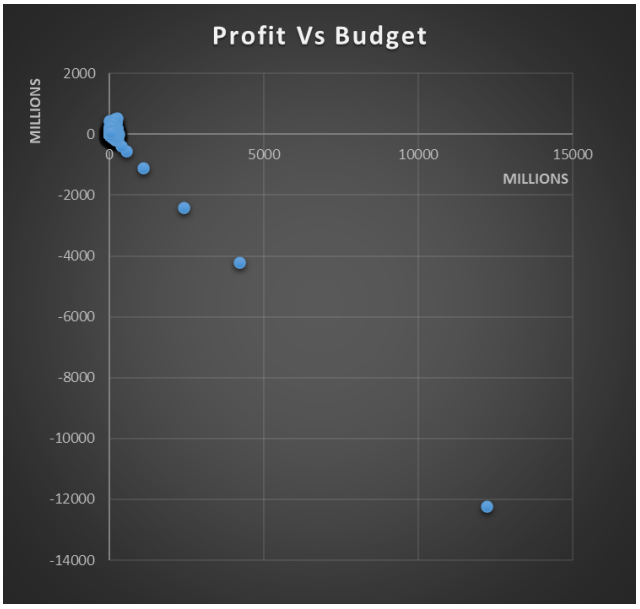
Before cleaning we had 5044 Columns(Including Title on first column) After cleaning we have 3744 Columns(Including Title on first column).

Cleaned Data: Data Cleaning

(B) Movies with the 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? (Movies with highest profit)



Outliers:

-12213298588
-4199788333
-2499804112
-2397701809
-2127109510

Top 5 Profitable Movies:

director_name	actor_1_name	movie_title	title_year	imdb_score	Profit
James Cameron	CCH Pounder	Avatar	2009	7.9	523505847
Colin Trevorrow	Bryce Dallas Howard	Jurassic World	2015	7	502177271
James Cameron	Leonardo DiCaprio	Titanic	1997	7.7	458672302
George Lucas	Harrison Ford	Star Wars: Episode IV - A New Hope	1977	8.7	449935665
Steven Spielberg	Henry Thomas	E.T. the Extra-Terrestrial	1982	7.9	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.

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.

Your task: Find IMDB Top 250 (IMDB TOP 250)

(D) Best Directors:

TGroup 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 top 10 directors. In case of a tie in IMDb score between two directors, sort them alphabetically.

Your task: Find the best directors

Best Directors: (Top 10 Directors)

Top 10 Directors	Average of imdb_score
Charles Chaplin	8.60
Tony Kaye	8.60
Alfred Hitchcock	8.50
Damien Chazelle	8.50
Majid Majidi	8.50
Ron Fricke	8.50
Sergio Leone	8.43
Christopher Nolan	8.43
Asghar Farhadi	8.40
Marius A. Markevicius	8.40

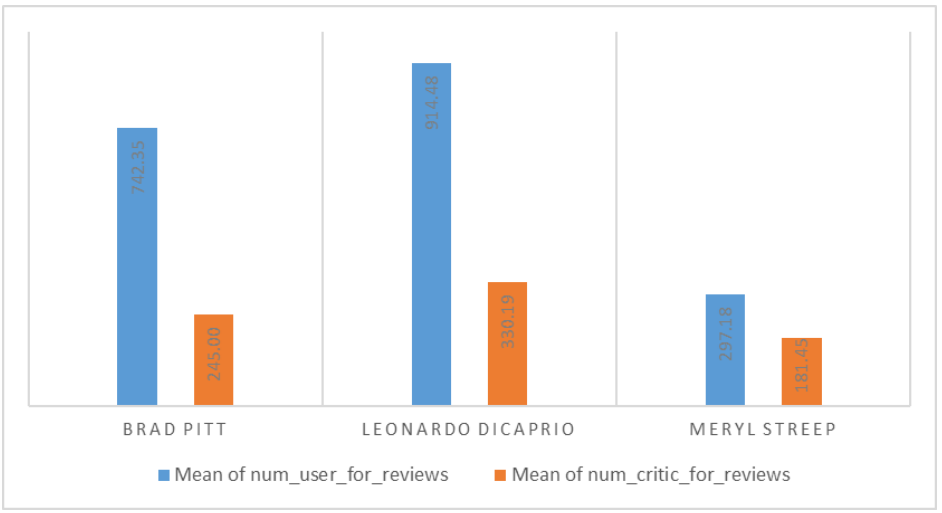
(E) Popular GENRES:

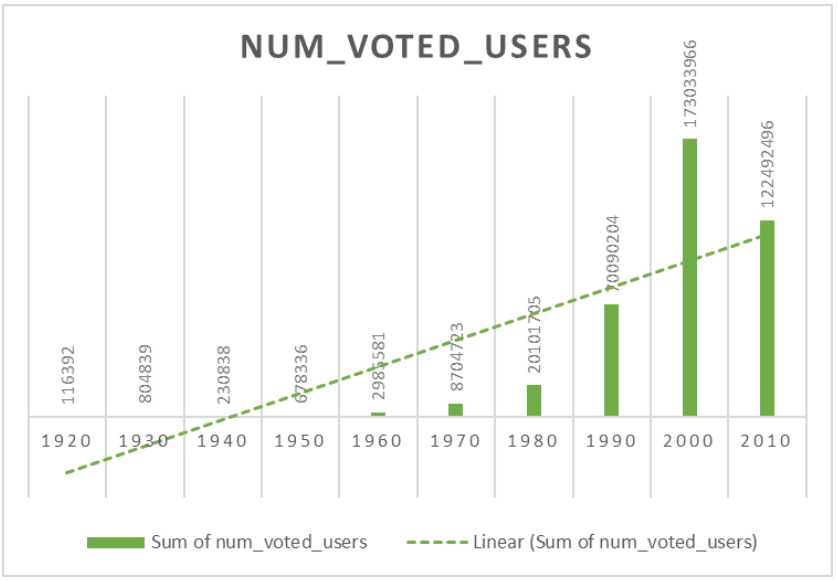
Perform this step using the knowledge gained while performing previous steps.

Your task: Find popular genres

Popular Genres: GENRES

(F) CHARTS:





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

actor_1_name	Mean of num_user_for_reviews	Mean of num_critic_for_reviews
Brad Pitt	742.35	245.00
Leonardo DiCaprio	914.48	330.19
Meryl Streep	297.18	181.45

RESULT

I gained proper skills and knowledge about MS-Excel and its formulas, Bar graphs, Charts, Pivot Table . Some sort of knowledge about Data Cleaning, and how we perform Data Analysis using MS-Excel.