IMDB Movie Analysis

Project description:

The Internet Movie Database (IMDb) is an online database containing information and statistics about movies, TV shows and video games as well as actors, directors and other film industry professionals. This information can include lists of cast and crew members, movie release dates and box office information, plot summaries, trailers, actor and director biographies and other trivia.

In this Project, a large set of raw data is given and asked to find the 5TA major tasks by doing data analysis after cleaning the given raw data

Approach:

First, I read the problem tasks asked to find and downloaded the raw data from the website. Raw data had many unwanted columns which are of no use to the analysis and it also had many null values and duplicates. So I cleaned the data and started doing analysis one by one in Microsoft excel using Tables, Conditional formatting, Sorting, filtering, Charts, Pivot Tables and pivot charts

Tech stack used:

Microsoft Excel 2021

Insights:

I learned more about Conditional formatting in this project, understood what it is really capable of when it solved imdb top 250 movies task. I also got handy to Pivot tables more and its easy to interpret the large data as it summarizes the data quickly. I got to understand how in real life data analysts solve such kind of problems by doing analysis. This project gave a real feel of analysing data

Results:

I retrieved the data from the given datasets by performing options and tools in Microsoft excel and the results obtained are mentioned as follows. These summarized results would help the movie producing company to take better data driven decisions

I have attached a drive link below containing the Excel sheet that I have worked **It contains solutions for all the tasks** and steps I followed while completing them have explained below

EXCEL SOLUTION File Google drive link

EXCEL SOLUTION File Microsoft One drive link

Since Google sheets don't have many features, I have shared One drive link as well

Task 1:

Cleaning the data

Data cleaning process I used:

1)First I Found columns which are of no use in the analysis like actor1 , actor 3 Facebook likes and then I dropped those columns

2)Then I Found the primary key column here (Movie title) I selected that particular column and Removed duplicates using conditional formatting and remove duplicates icon from Data tab

3)After that I found blanks by clicking Find & Select option from the Home tab and deleted the entire row which contains blanks

Before cleaning (Raw data)	After cleaning
5044 rows	3773 rows
27 columns	18 columns

1.Columns dropped:

1) actor_3_facebook_likes
2) actor_1_facebook_likes
3) actor_2_facebook_likes
4) cast_total_facebook_likes
5) facenumber_in_poster
6) plot_keywords
7) movie_imdb_link
8) content_rating
9) aspect ratio

2.Duplicates:

126 duplicates were found in the Movie title column and they were removed from that column

3.Null/Blank Values:

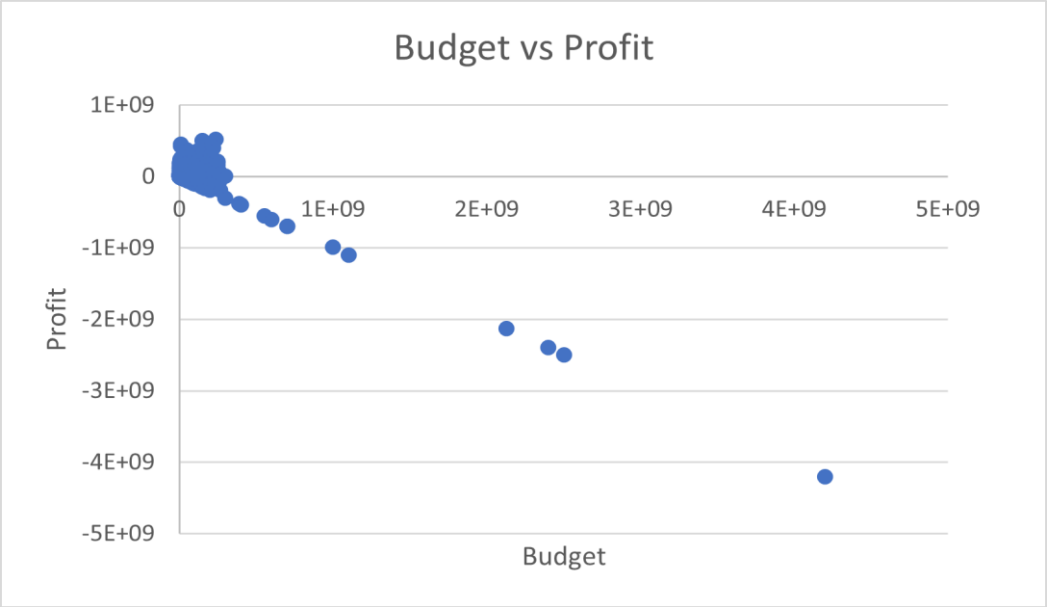
1145 blank values were found and they were removed

Task 2:

Movies with highest profit

I created a new column Profit (Gross-Budget) and sorted it to find out the top 10 movies with highest profit.

1. AvatarÂ
2. Jurassic WorldÂ
3. TitanicÂ
4. Star Wars: Episode IV - A New HopeÂ
5. E.T. the Extra-TerrestrialÂ
6. The AvengersÂ
7. The Lion KingÂ
8. Star Wars: Episode I - The Phantom MenaceÂ
9. The Dark KnightÂ
10. The Hunger GamesÂ



Using scatter charts, I have found Outliers in budget vs profit graph

Most of the data points are clustered at one location while few are located away from that denser region of points

Task 3

Top 250:

Using Conditional formatting option from the home tab, I chose a set of icons from icon sets, in which I added a new rule (Green icon means the num_of_voted_users = > 25000 and red icon means num_of_voted_users <24999)

Then I used Multi level sort option from the Data tab and sorted the num of voted users by Green and sorted the imdb score from largest to smallest

Then i obtained a result of top movies, then i created a new column **imdb top 250** and copy pasted 250 movies from **Movie title** column

Created a new column **RANK** and gave rank for the top movies

Using Filter option in the **Language** column, filtered out all the Non English movies which is in imdb top 250 and extracted the data and stored in the new column **Top foreign Lang Movies**

Task 4:

Best Directors

For this task, I created a pivot table from the data table and value filtered top 10 best directors based on their mean imdb score

And using this data, I added a new column **Top 10 directors** in main data table

Top 10 Directors	
Director Names	**Average of imdb_score**
Alfred Hitchcock	8.5
Asghar Farhadi	8.4
Charles Chaplin	8.6
Christopher Nolan	8.425
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
Sergio Leone	8.433333333
Tony Kaye	8.6
Grand Total	**8.45**

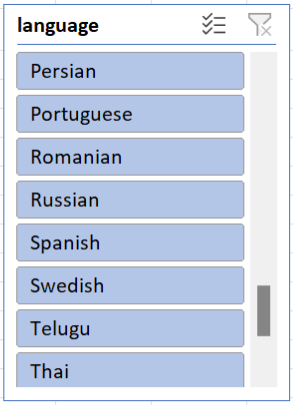
Task 5:

Popular genres

I created pivot table from the table containing row labels as Genres and thought num of user reviews and critic reviews decide the popularity factor

Also inserted a slicer of languages

Top 10 Popular Genres		
Genres	**Sum of num_critic_for_reviews**	**Sum of num_user_for_reviews**
Action|Adventure|Sci-Fi	17915	43207
Action|Adventure|Thriller	10207	25195
Comedy	16968	33166
Comedy|Drama	17497	25871
Comedy|Drama|Romance	20369	32763
Comedy|Romance	16340	25659
Crime|Drama|Thriller	14873	28125
Drama	20786	42069
Drama|Romance	16998	33925
Horror	10167	23342
Grand Total	**162120**	**313322**



Task 6:

Charts

I created 3 new columns **Meryl sheep, Brad pitt, leo_Caprio** and inserted the movies of these three actors using the **actor 1 name** column (filtered and extracted)

Then I created a new column **Combined** and added all the three actor movies in a single column

I created a pivot table to find out Critic favourite and audience favourite actors

From the pivot table, I inferred that **Leonardo Di caprio** has the highest mean

Actors	**Average of num_critic_for_reviews**	**Average of num_user_for_reviews**
Brad Pitt	245	742.3529412
Leonardo DiCaprio	322.2	922.55
Meryl Streep	181.4545455	297.1818182
Grand Total	**262.6041667**	**715.4166667**

To find out the most number of voted users by decade,

First I grouped the years manually by decade using **title year** column

Then using pivot table, I found that during 2000s there were large number of voted users

Decade	Sum of num_voted_users
1920s	116387
1930s	804839
1940s	159517
1950s	678336
1960s	2983442
1970s	8269031
1980s	19344372
1990s	69635866
2000s	165946388
2010s	116072506
Grand Total	**384010684**

Pivot chart (Bar graph):