Name: **AASHIK BHARATH EGNYA VARAHAN**
Student No.: **R00182866**

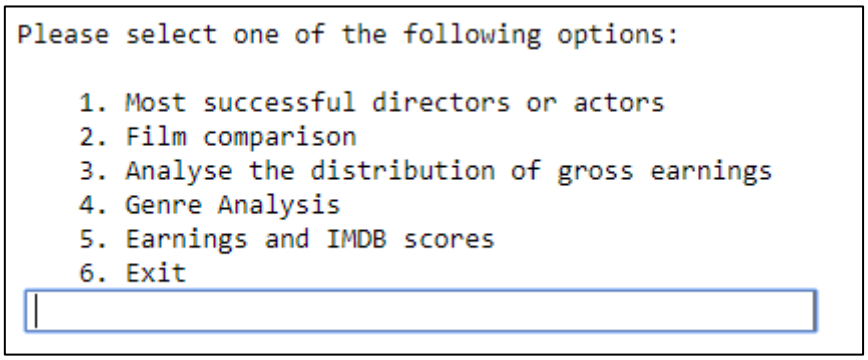
ANALYTICAL PROGRAMMING PROJECT

INTRODUCTION:

The objective of this project is to create an application which analyses the given IMDB movie dataset and explore some of the interesting aspects in it. The IMDB movie dataset hosted by Kaggle contains the details of over 5043 movies scraped from the IMDB movie website. The dataset attempted to collect 28 features describing each movie such as Movie Title, Color, Critic Reviews, Duration of movie, Actors names, Director Names, Gross revenue, Genre, Language, Country, Content Rating, Budget, Title Year, IMDB Score and Aspect Ratio. The dataset contains some missing values and these missing values have been removed from the dataset.

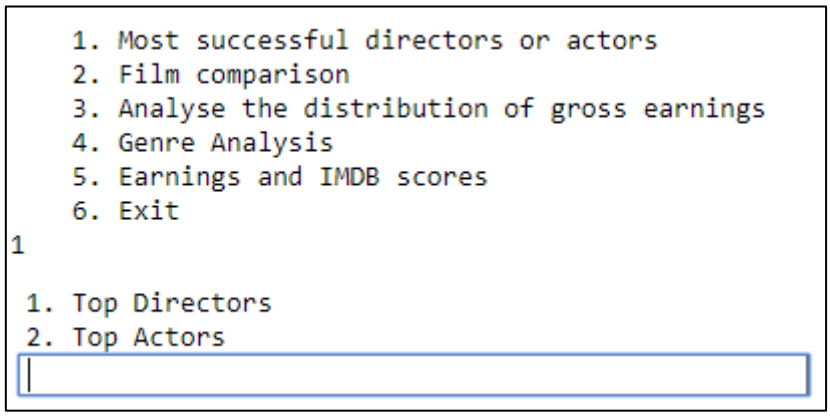
The application gives the user a set of options to perform analysis on the dataset. When the application ***main()*** function is called, the user is displayed with below options.

The application uses inbuilt packages such as **pandas** and **numpy** to analyse the data, **matplotlib** and **seaborn** to visualize and plot the analysed data.



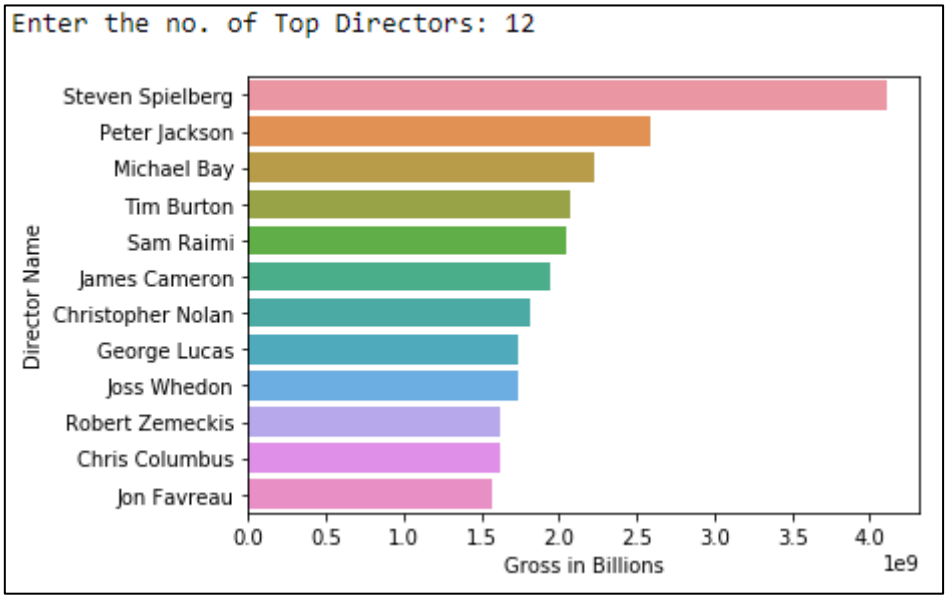
Menu Option 1 – Most successful directors or actors:

On selecting the first menu option, the user gets two options to select from, either Top Directors or Top Actors.

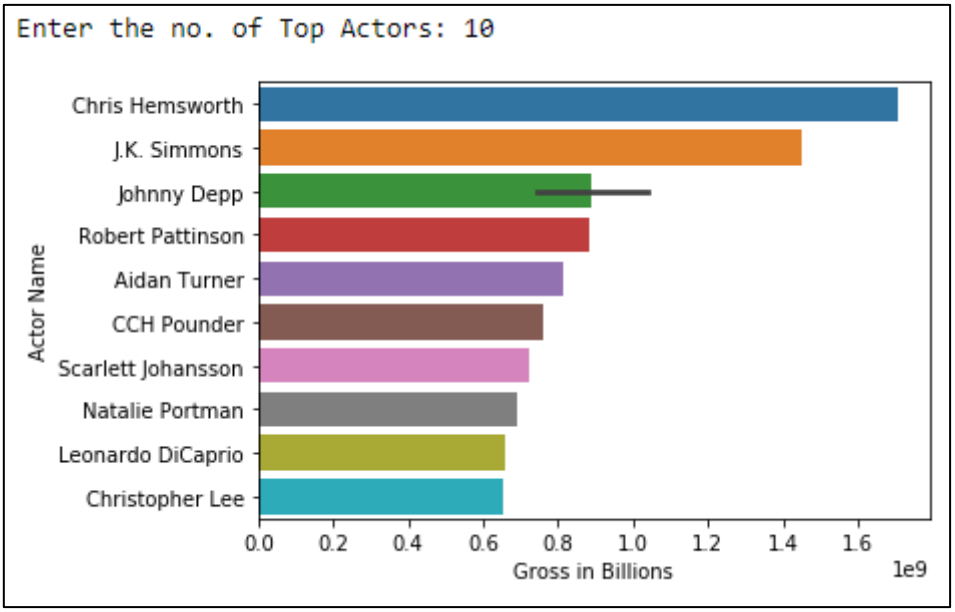


If the user selects the first option Top Directors, the user is again requested with an option to enter the number of directors to be displayed based on the gross film earnings.

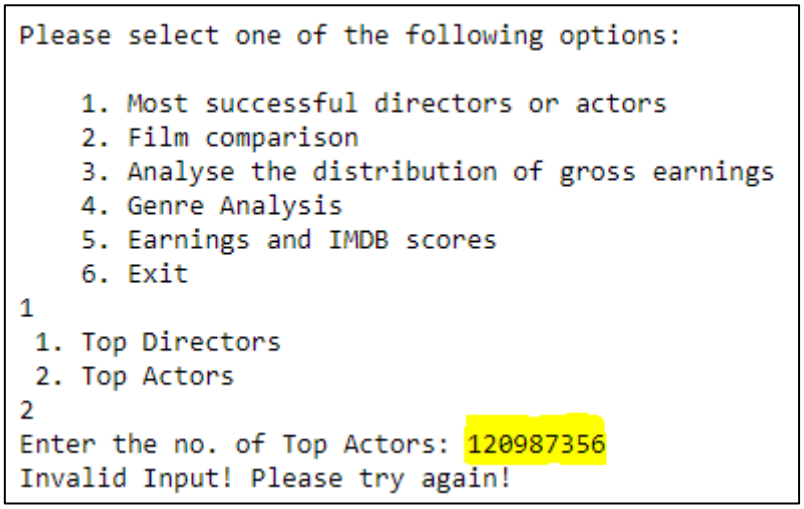
For example, if the user gives an input of 12 then a list of top 12 directors based on gross film earnings is displayed as a horizontal bar graph.



If the user selects the second option Top Actors, the user is prompted with an option to enter the number of actors to be displayed based on gross film earnings. For example, if the user gives an input of 10 then a list of top 10 actors based on gross film earnings is displayed as a horizontal bar graph.



The application deals with the basic error checking on director and actors names. It checks the value given by the user, i.e., if the number input is negative or if the number input is greater than the actual number of directors/actors, the user is displayed with an error message and again prompted with the option to enter the number of actors/directors. This process is repeated until the user enters a valid input.

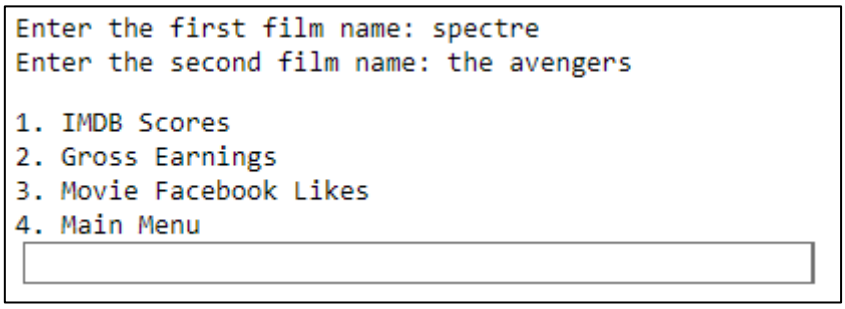


Menu Option 2 – Film Comparison:

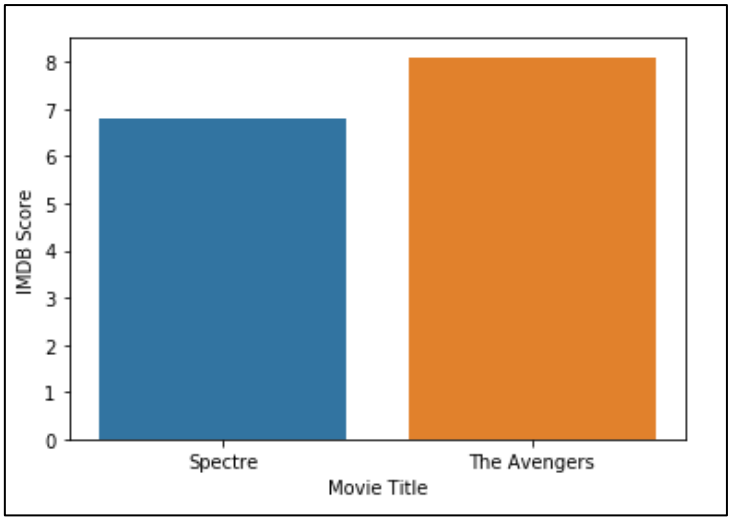
The second option in the menu performs film comparison. The user is asked for two film names when this menu option is selected and the applications checks the validity of both the inputs of film names given.

If the movie names entered are invalid, the user will again be requested to enter the movie names. The application repeats asking the user to enter two movie names till they enter a valid film name.

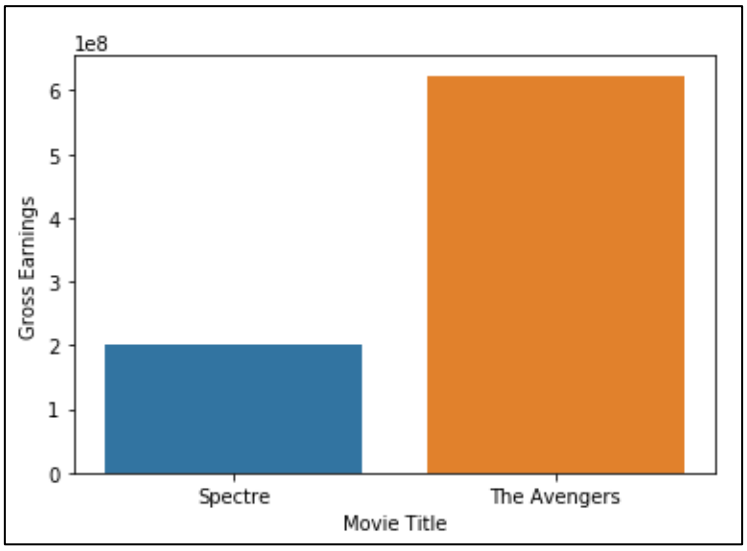
If two film names entered are valid, the user is given with a set of options like below:



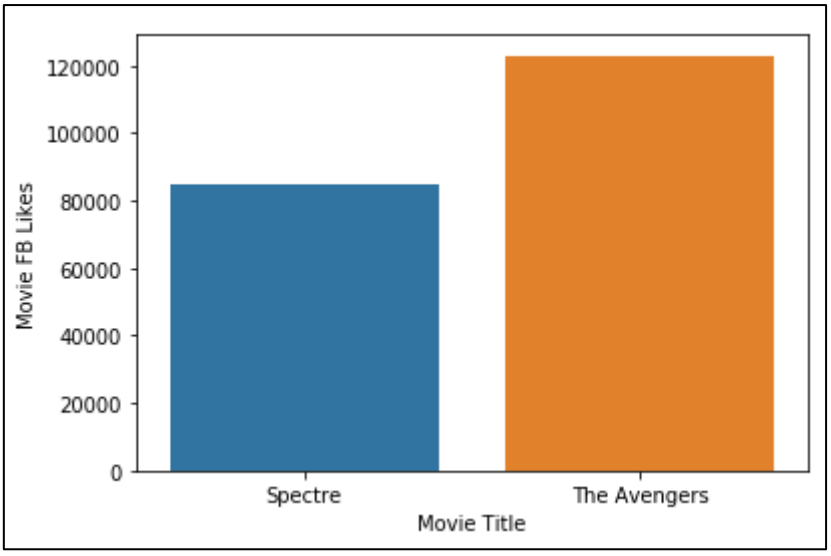
When the user selects the first option, a bar graph is generated against the two film names given as input and their respective IMDB scores.



If the user selects the second option, a bar graph for total gross earnings of two film names are generated.



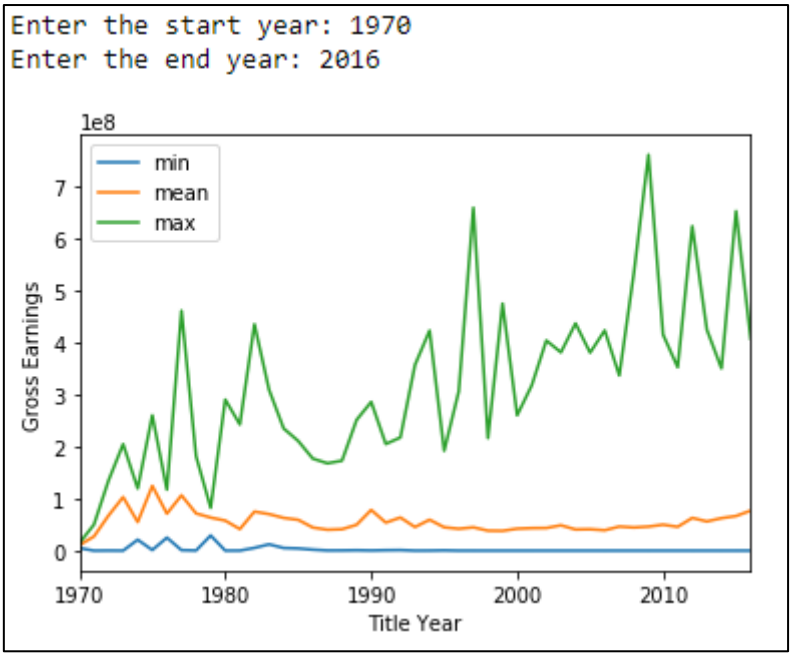
If the user selects the third option, a bar graph for total Facebook likes of both the film names are generated.



If the user selects the last option, the user is again taken back to the main menu options.

Menu Option 3 - Analyse the distribution of gross earnings

When this option is selected, the user is asked to input the start and end year for the analysis of gross earnings. The application does basic error checking on the input by checking whether the year entered is valid and within the range of dataset. If both the years entered are valid, the application subsets the data by **groupby** and **aggregate** functions. Then, a line plot is generated for the subset data against three different values min(minimum), mean(average) and max(maximum) values of gross earnings for all the years between the two given inputs (inclusive of the years given as input).



If the user enters invalid year input, an error message is displayed and the user is again prompted to enter two valid year inputs.

Menu Option 4 – Genre Analysis

When the user selects this option, the program displays a list of unique genres in the dataset. The user is given an option to enter a genre from the displayed list. The application then displays the mean IMDB Score of all the films within that genre.

The mean IMDB score is calculated by checking whether the user given genre is available in the unique list. If it is available, a subset data with indices of the selected genre is generated where available. The mean IMDB Score is then calculated for that particular subset data.

If the user choice is not in the genre list displayed, then an error message is displayed and the user will be repeatedly prompted to enter a genre name until a valid genre is entered.

The below list is generated when the 4th option is selected:

```
List of Genres Analysis:

Action
Adventure
Animation
Biography
Comedy
Crime
Documentary
Drama
Family
Fantasy
Film-Noir
Game-Show
History
Horror
Music
Musical
Mystery
News
Reality-TV
Romance
Sci-Fi
Short
Sport
Thriller
War
Western
```

The user selects any one genre from the above list. Here, we have entered Romance as genre.

```
Enter a Genre for IMDB Score: Romance
IMDB score for Romance is: 6.45
```

Menu Option 5 – Earnings and IMDB scores

When the user selects this option, a relationship between the numerical IMDB scores and other numerical columns in the dataset is found to help the user build a model based on the correlation.

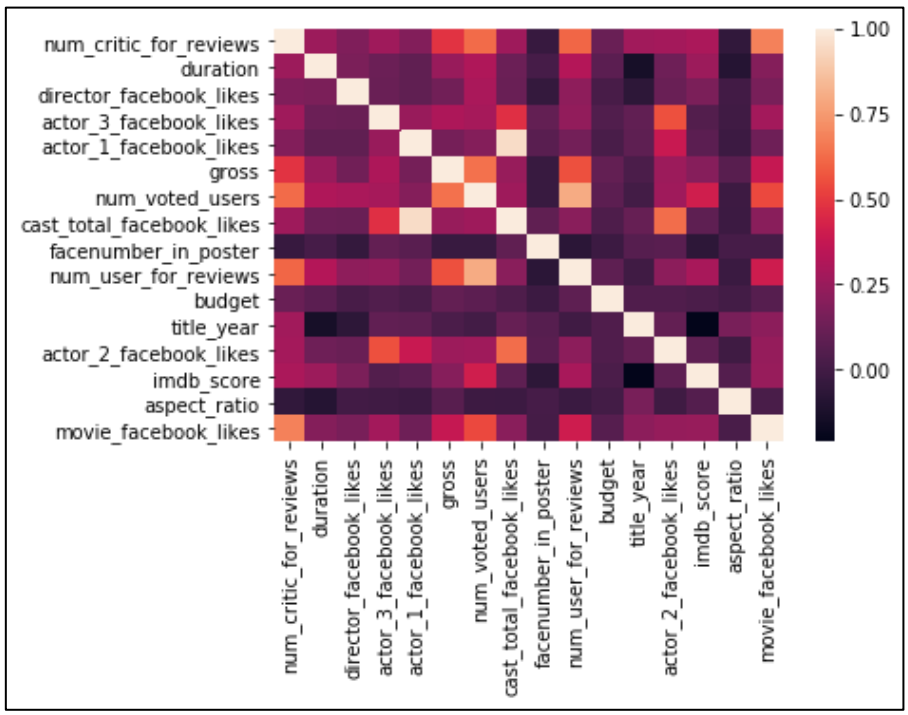
The below table shows the correlation between IMDB Score and other numerical values:

```
Correlation btween imdb_score and other numerical variables:

num_critic_for_reviews       0.305303
duration                     0.261662
director_facebook_likes      0.170802
actor_3_facebook_likes       0.052633
actor_1_facebook_likes       0.076099
gross                        0.198021
num_voted_users              0.410965
cast_total_facebook_likes    0.085787
facenumber_in_poster        -0.062958
num_user_for_reviews         0.292475
budget                       0.030688
title_year                  -0.209167
actor_2_facebook_likes       0.083808
imdb_score                   1.000000
aspect_ratio                 0.059445
movie_facebook_likes         0.247049
Name: imdb_score, dtype: float64
```

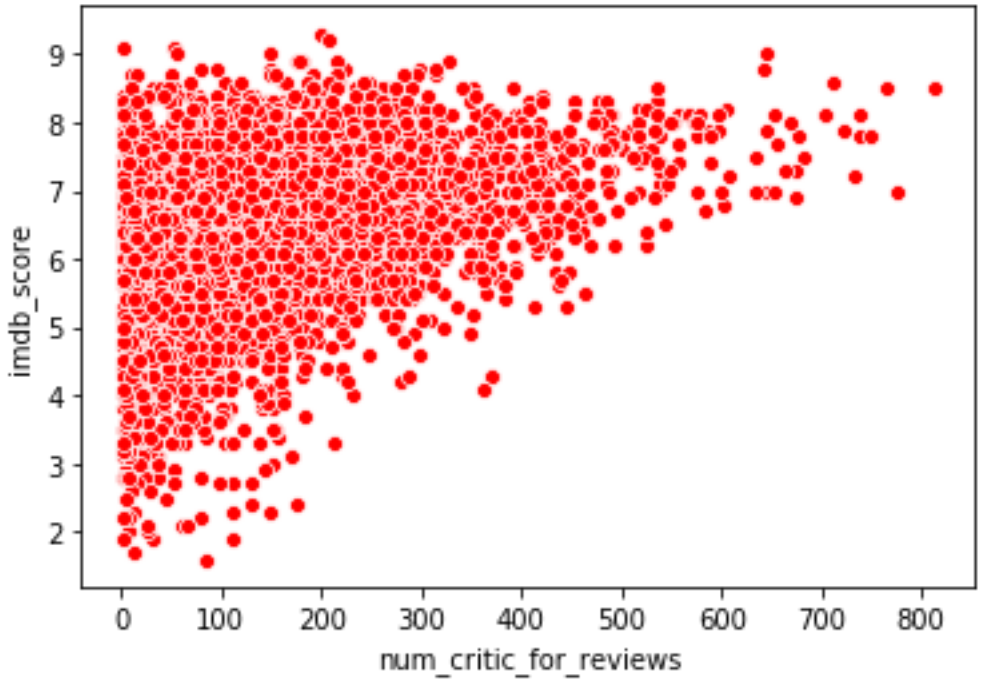
From the above table, we can infer that **num_voted_users** has the highest correlation relation against IMDB Score of 0.41 followed by **num_critic_for_reviews** with correlation of 0.31

The below heatmap plot shoes the correlation relation:



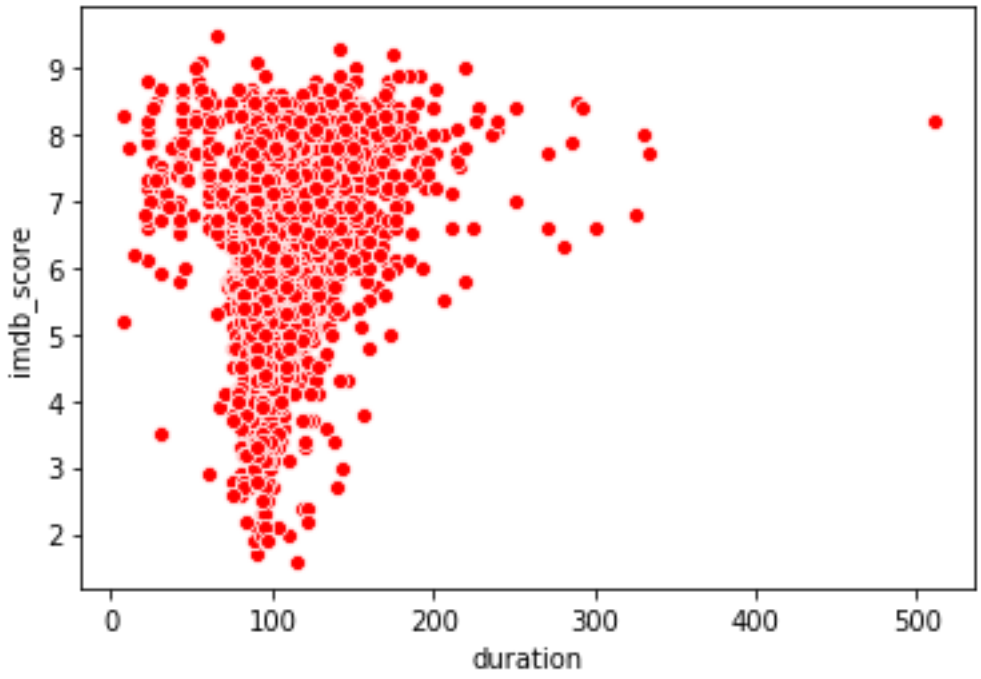
Now, we will find the relationship between each of the numerical column and IMDB Scores.

Relationship between num_critic_for_reviews and imdb_score:



The correlation coefficient between num_critic_for_reviews and imdb_score is 0.305
A weak positive relationship exists between these two.

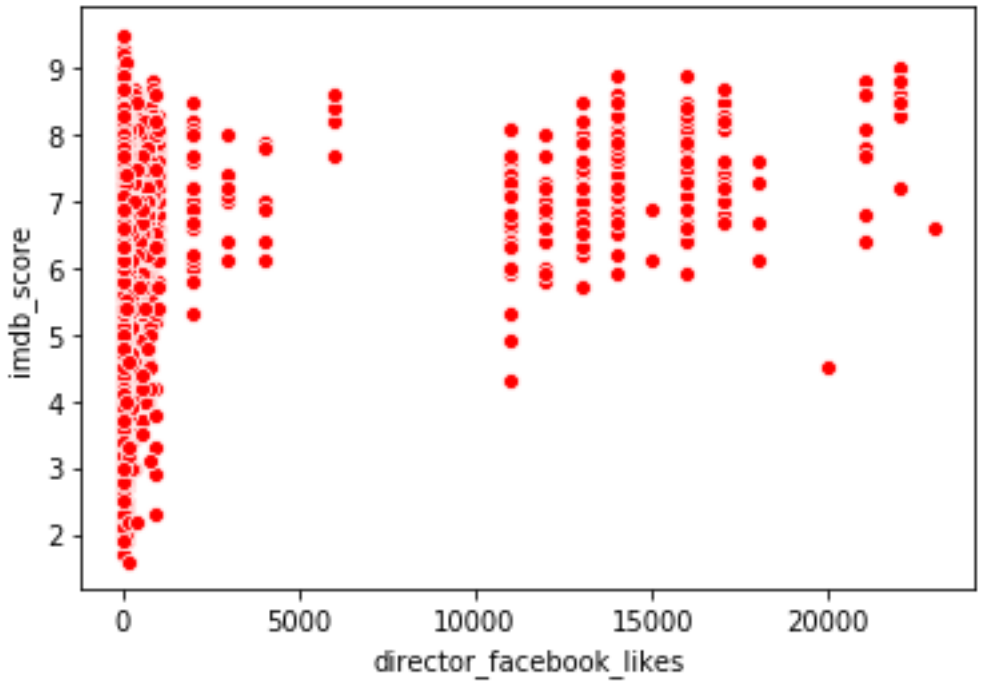
Relationship between duration and imdb_score:



The correlation coefficient between duration and imdb_score is 0.262

A weak positive relationship exists between these two.

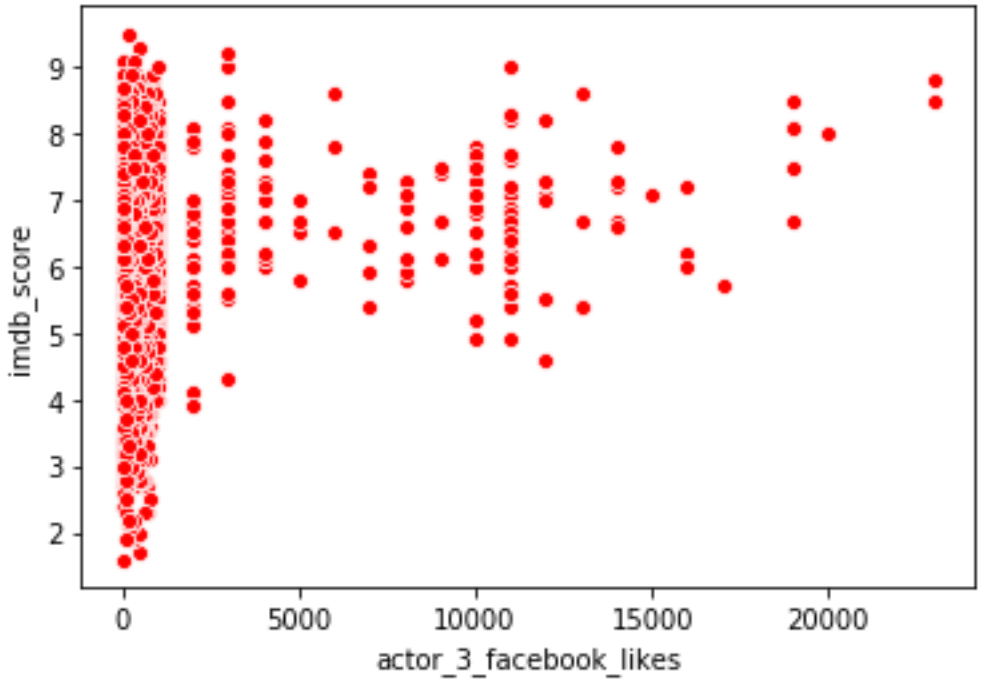
Relationship between director_facebook_likes **and imdb_score:**



The correlation coefficient between director_facebook_likes and imdb_score is 0.171

A weak positive relationship exists between these two.

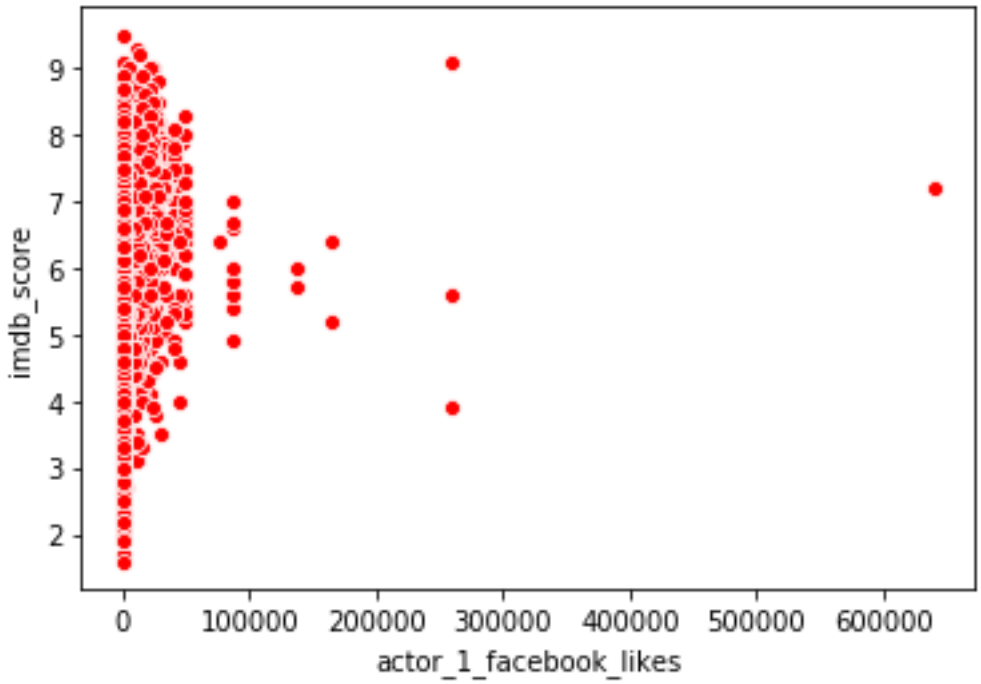
Relationship between actor_3_facebook_likes and imdb_score:



The correlation coefficient between actor_3_facebook_likes and imdb_score is 0.053

Almost, No correlation relationship exists between these two.

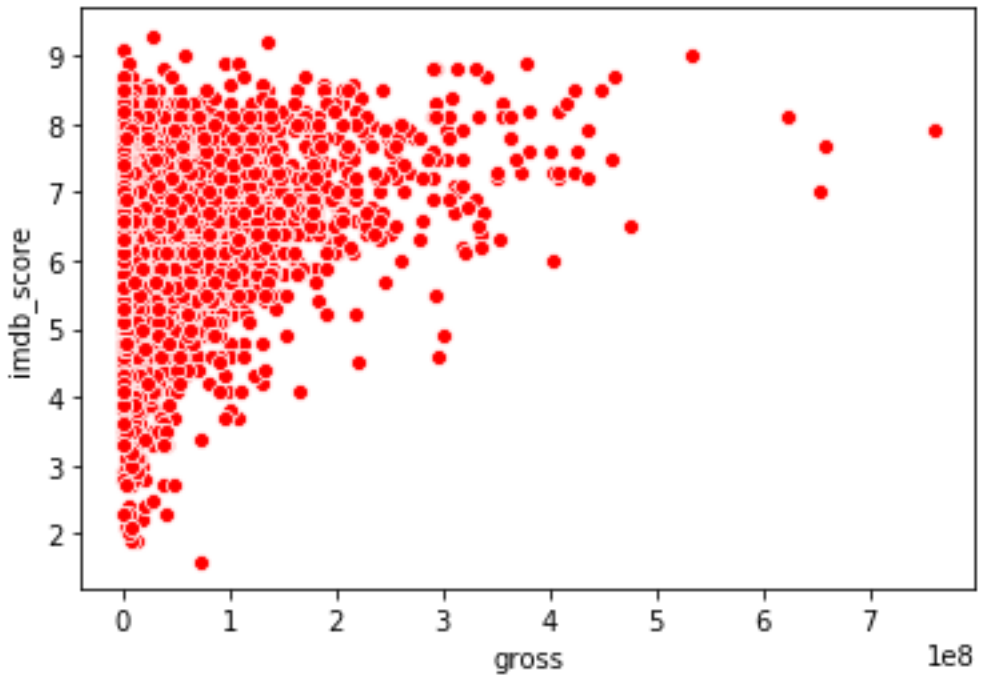
Relationship between actor_1_facebook_likes and imdb_score:



The correlation coefficient between actor_1_facebook_likes and imdb_score is 0.076

Almost, No correlation relationship exists between these two.

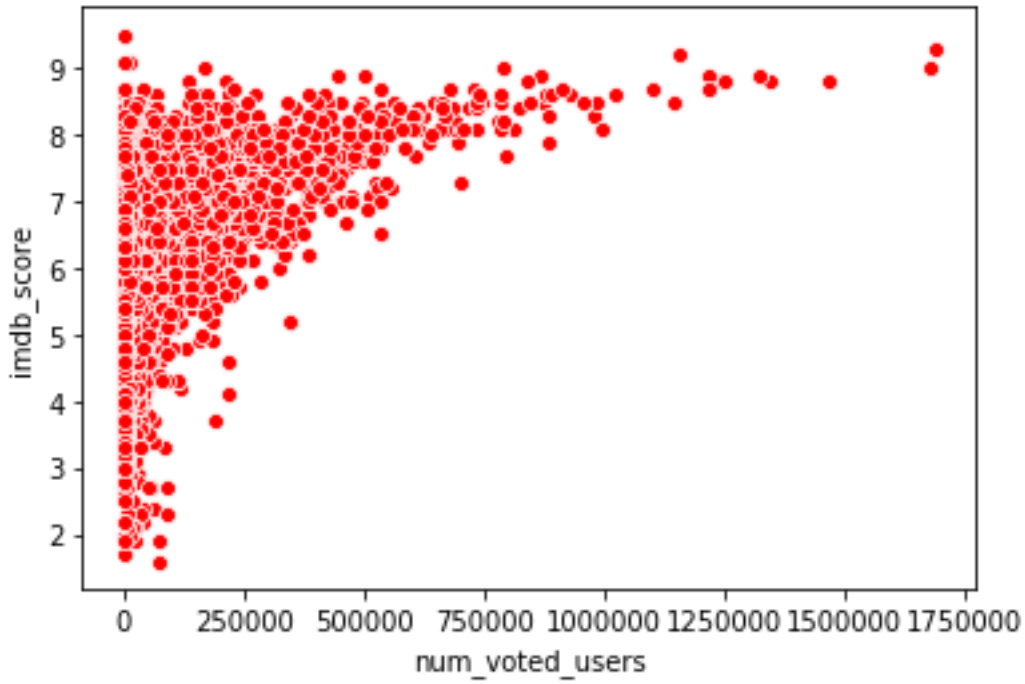
Relationship between gross earnings and imdb_score:



The correlation coefficient between gross and imdb_score is 0.198

A weak positive correlation relationship exists between these two.

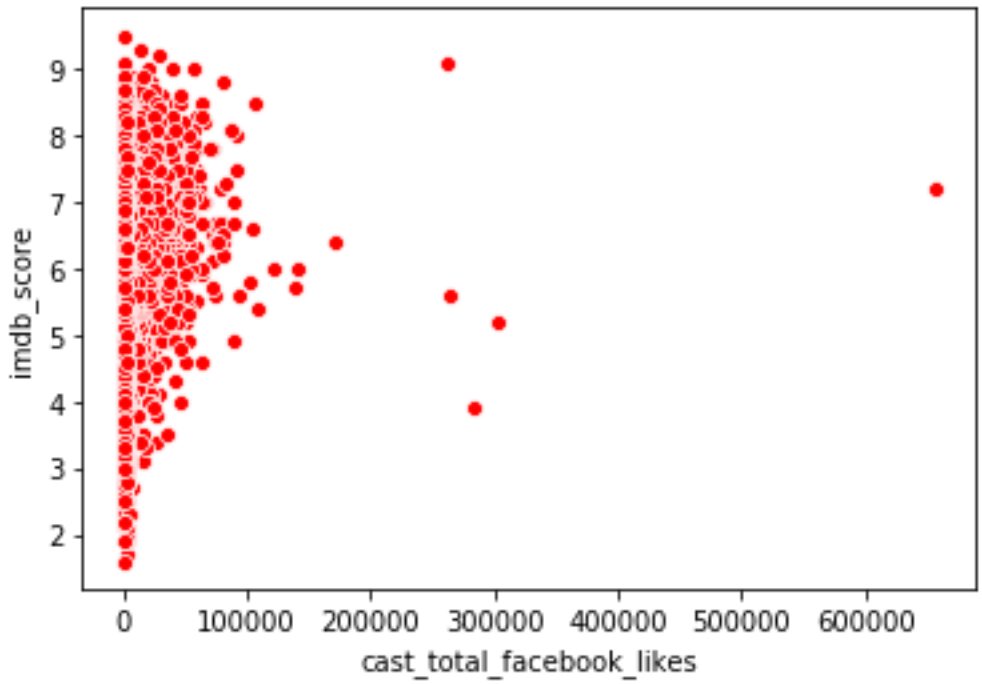
Relationship between num_voted_users and imdb_score:



The correlation coefficient between num_voted_users and imdb_score is 0.411

A weak positive correlation relationship exists between these two.

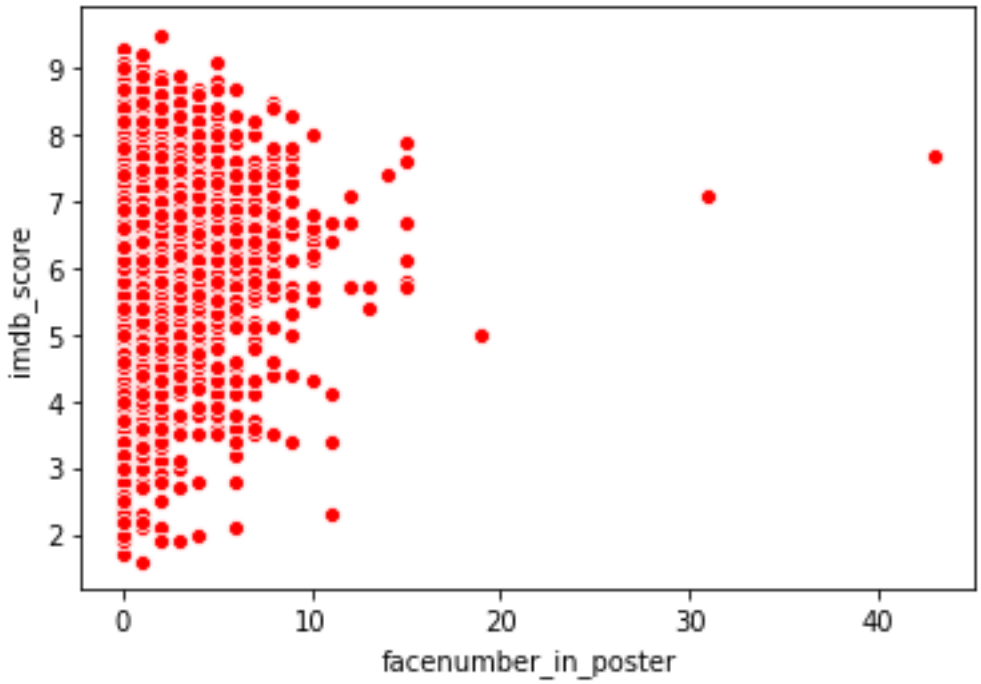
Relationship between cast_total_facebook_likes and imdb_score:



The correlation coefficient between cast_total_facebook_likes and imdb_score is 0.086

Almost, No correlation relationship exists between these two.

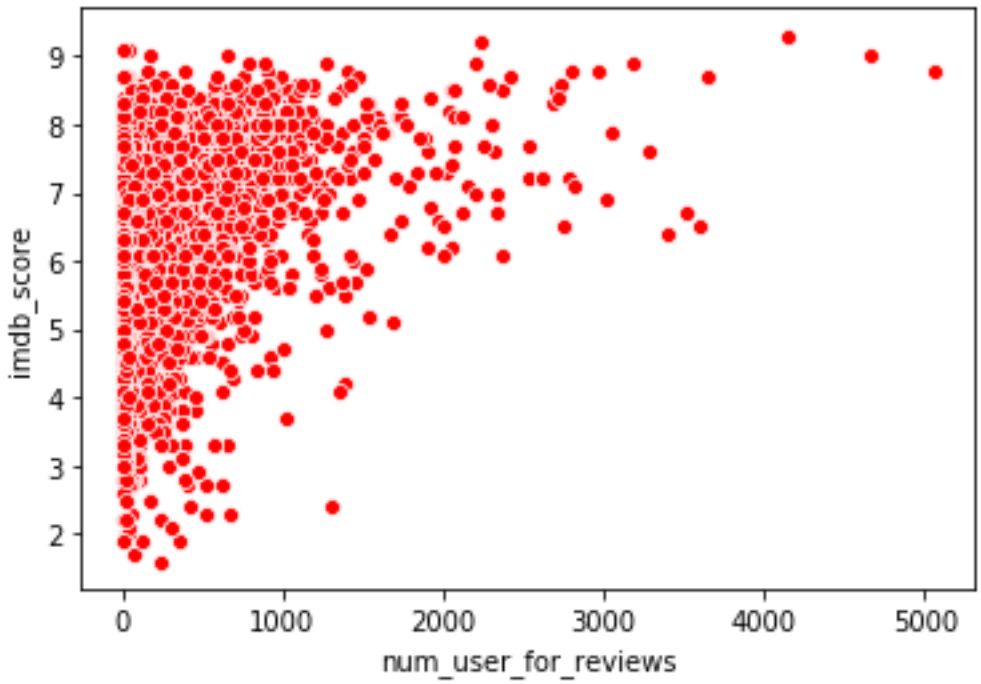
Relationship between facenumber_in_poster and imdb_score:



The correlation coefficient between facenumber_in_poster and imdb_score is -0.063

A very weak negative correlation relationship exists between these two.

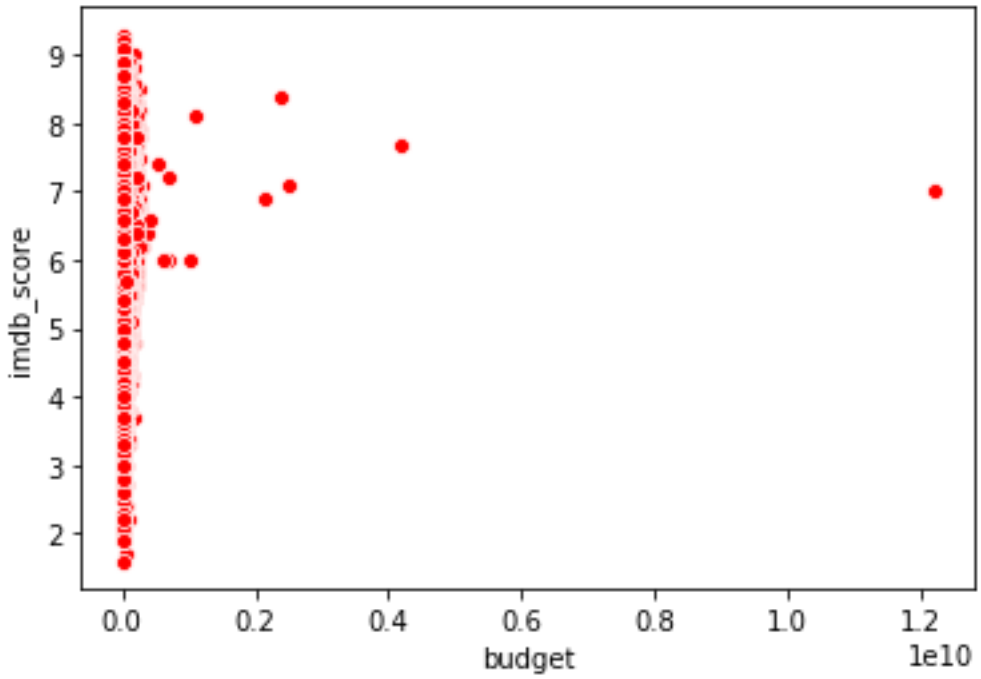
Relationship between num_user_for_reviews and imdb_score:



The correlation coefficient between num_user_for_reviews and imdb_score is 0.292

A weak positive correlation relationship exists between these two.

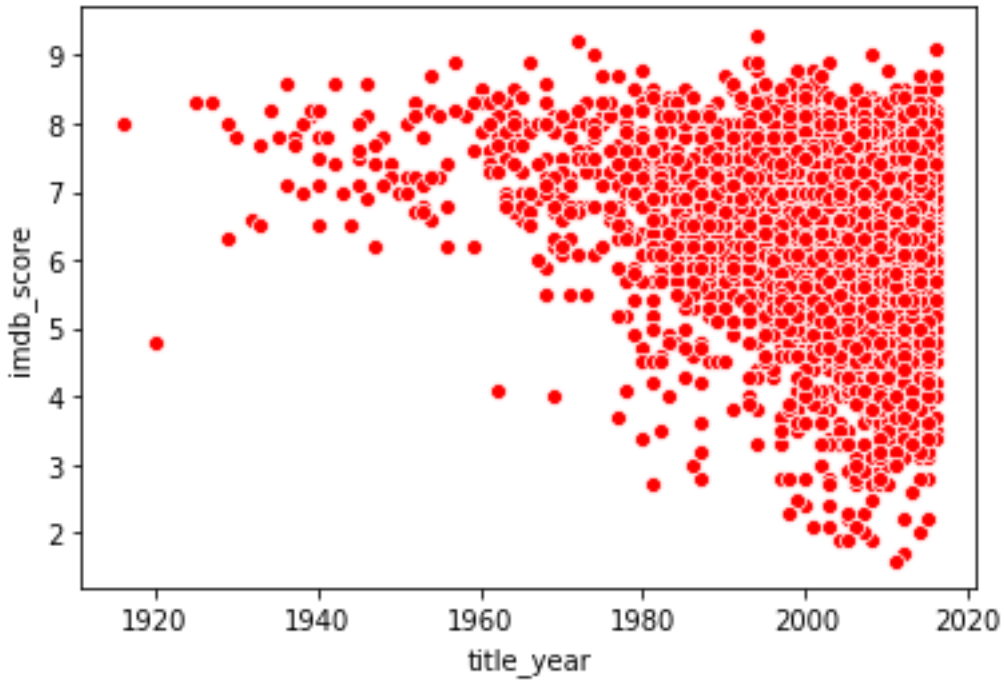
Relationship between budget and imdb_score:



The correlation coefficient between bdget and imdb_score is 0.03

Almost, no correlation relationship exists between these two.

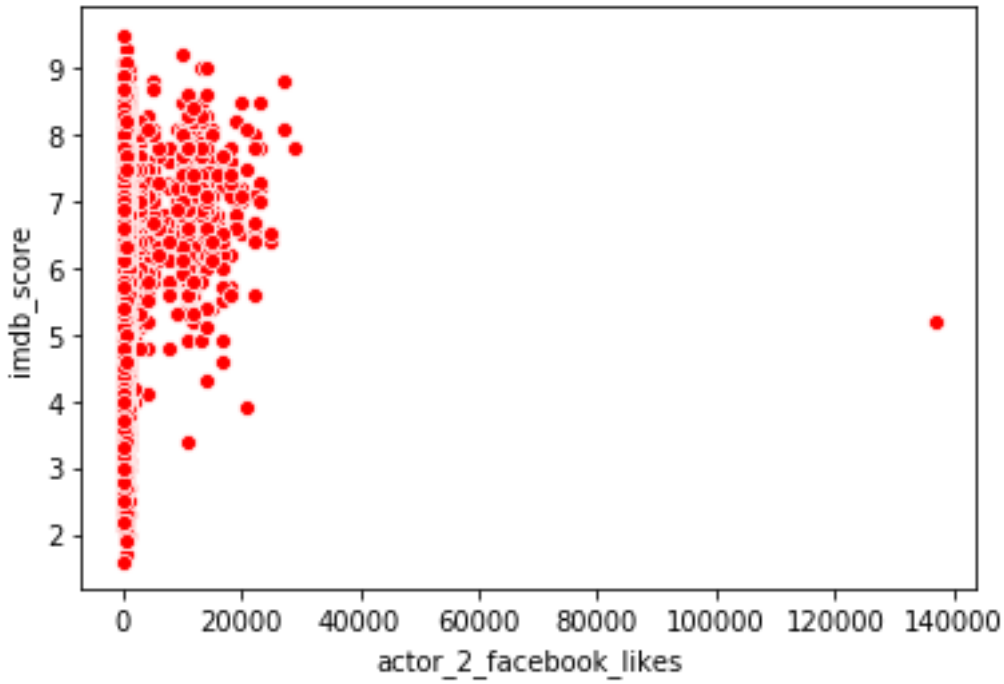
Relationship between title_year and imdb_score:



The correlation coefficient between budget and imdb_score is -0.209

A weak negative correlation relationship exists between these two.

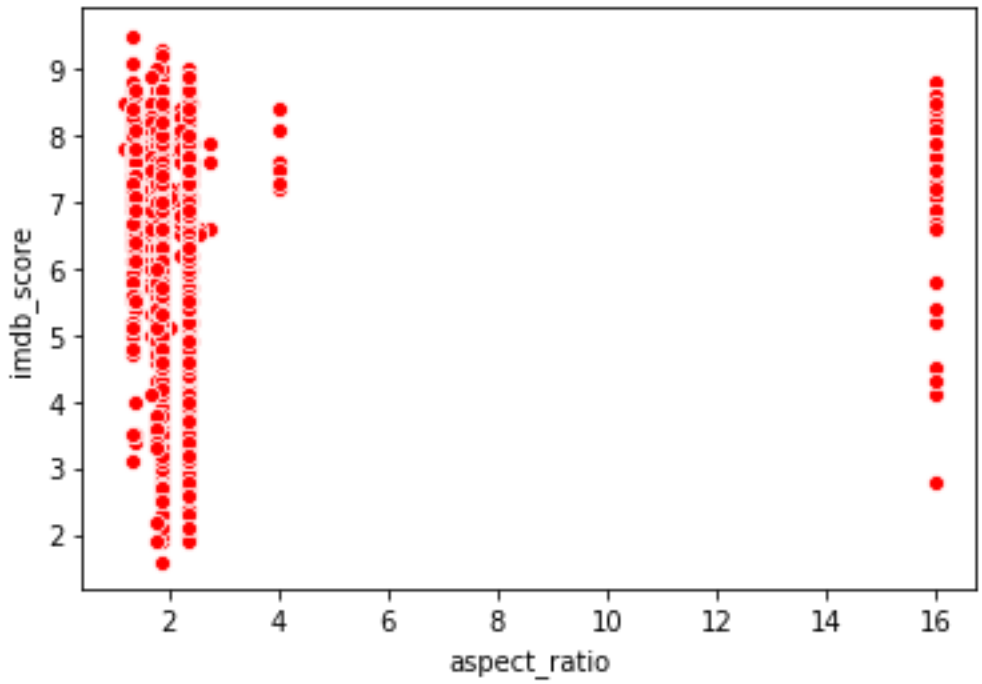
Relationship between actor_2_facebook_likes and imdb_score:



The correlation coefficient between actor_2_facebook_likes and imdb_score is 0.083

A weak positive correlation relationship exists between these two.

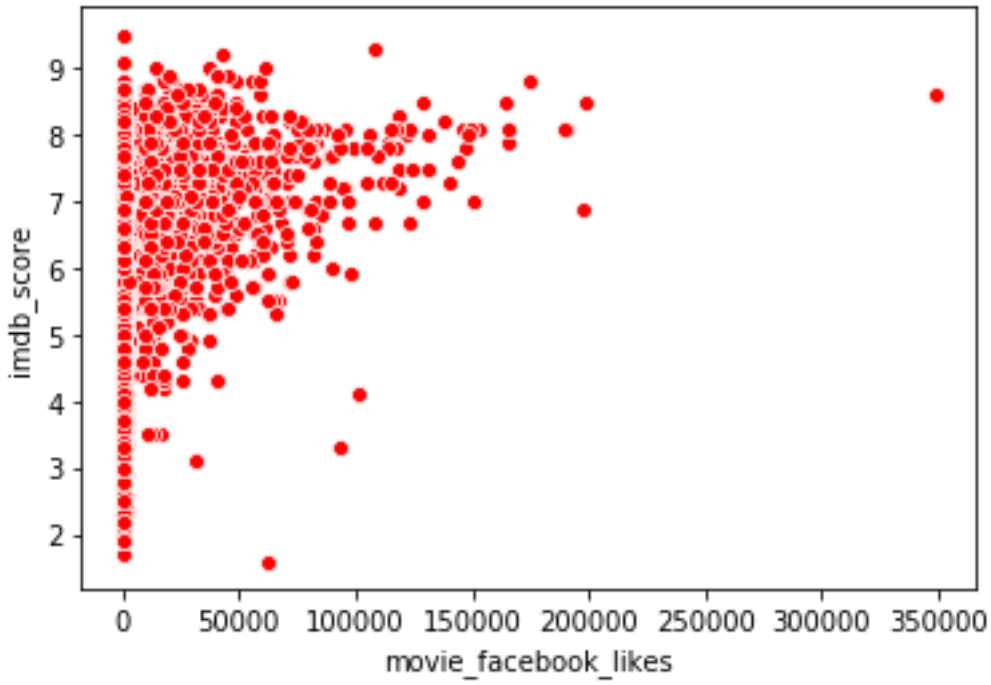
Relationship between aspect_ratio and imdb_score:



The correlation coefficient between aspect_ratio and imdb_score is 0.059

A very weak positive correlation relationship exists between these two.

Relationship between movie_facebook_likes and imdb_score:



The correlation coefficient between aspect_ratio and imdb_score is 0.247

A weak positive correlation relationship exists between these two.

From the above plots, we can find that correlation between **num_voted_users** and **imdb_score** is high comparatively to the relation between other numerical variables and imdb_score. This suggests to select the num_voted_users as the primary explanatory variable to build the model that predicts IMDB score of each film with a good level of accuracy and very significant.

Menu Option 6 – Exit

The application display the associated output when the user selects an option 1-5 and again comes back to the main menu again. The application exits and terminates the operations once the user selects the option 6 – Exit.

```
Please select one of the following options:

    1. Most successful directors or actors
    2. Film comparison
    3. Analyse the distribution of gross earnings
    4. Genre Analysis
    5. Earnings and IMDB scores
    6. Exit
6
Thanks for using IMDB Review System!
```

CONCLUSION:

The application is designed as per the given requirements and executes well in order. It analyses the given dataset and produces visualizations of the analysed data in the form of bar graphs, scatterplots, correlation plots and as table outputs. The application also finds the relationship between different variables and helps the user understand the dataset to build a better model and predict the response variable with higher accuracy.