**SCIENTIFIC PROG IN PYTHON ASSIGNMENT**

**STUDENT NAME: SHIVAANI KATRAGADDA**

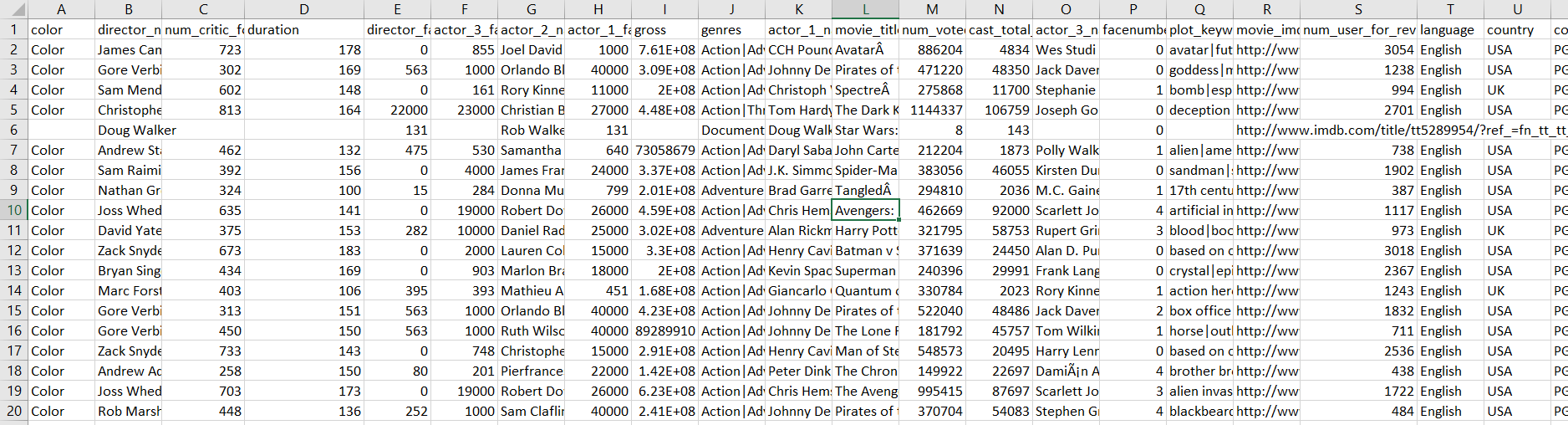
**STUDENT NUMBER: R00183214**

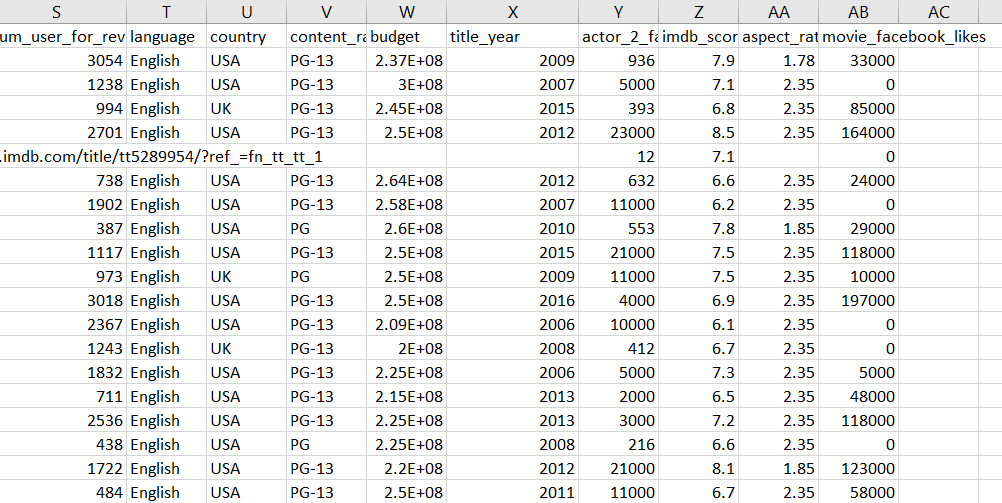
**REPORT**

In this assignment, I was given a movie\_metadata Excel file. It contains 28 columns and 5043 rows.

The 28 columns are as follows

|  |  |
| --- | --- |
| Column Name | Small Description of column |
| movie\_title | Title of the Movie |
| duration | Duration in minutes |
| director\_name | Name of the Director of the Movie |
| director\_facebook\_likes | Number of likes of the Director on his Facebook Page |
| actor\_1\_name | Primary actor starring in the movie |
| actor\_1\_facebook\_likes | Number of likes of the Actor\_1 on his/her Facebook Page |
| actor\_2\_name | Other actor starring in the movie |
| actor\_2\_facebook\_likes | Number of likes of the Actor\_2 on his/her Facebook Page |
| actor\_3\_name | Other actor starring in the movie |
| actor\_3\_facebook\_likes | Number of likes of the Actor\_3 on his/her Facebook Page |
| num\_user\_for\_reviews | Number of users who gave a review |
| num\_critic\_for\_reviews | Number of critical reviews on imdb |
| num\_voted\_users | Number of people who voted for the movie |
| cast\_total\_facebook\_likes | Total number of facebook likes of the entire cast of the movie |
| movie\_facebook\_likes | Number of Facebook likes in the movie page |
| plot\_keywords | Keywords describing the movie plot |
| facenumber\_in\_poster | Number of the actor who featured in the movie poster |
| color | Film colorization. ‘Black and White’ or ‘Color’ |
| genres | Film categorization like ‘Animation’, ‘Comedy’, ‘Romance’, ‘Horror’, ‘Sci-Fi’, ‘Action’, ‘Family’ |
| title\_year | The year in which the movie is released (1916:2016) |
| language | English, Arabic, Chinese, French, German, Danish, Italian, Japanese, etc |
| country | A country where the movie is produced |
| content\_rating | Content rating of the movie |
| aspect\_ratio | Aspect ratio the movie was made in |
| movie\_imdb\_link | IMDB link of the movie |
| gross | Gross earnings of the movie in Dollars |
| budget | The budget of the movie in Dollars |
| imdb\_score | IMDB Score of the movie on IMDB |

****

****

This is how the data looks like, from the two sheets, it is very clear that there is some data missing.

The libraries used in this assignment to do the tasks are as follows:

1.NUMPY:

Numpy package is a general-purpose array-processing package

2.PANDAS:

pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with structured (tabular, multidimensional, potentially heterogeneous) and time-series data both easy and intuitive.

3.PANDAS SERIES:

Pandas Series is a one-dimensional labeled array capable of holding data of any type (integer, string, float, python objects, etc.). The axis labels are collectively called index

4.MATPLOTLIB.PYPLOT:

matplotlib. pyplot is a collection of command style functions that make matplotlib work like MATLAB. Each pyplot function makes some changes to a figure.

5.SEABORN:

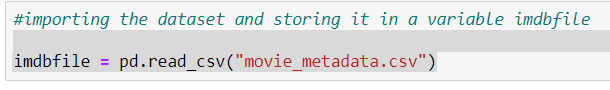
seaborn: statistical data visualization. Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

6.PANDAS\_PROFILING:

The pandas-profiling Python package is a great tool to create HTML profiling reports. For a given dataset, it computes the following statistics: Essentials: type, unique values, missing values. Quantile statistics like minimum value, Q1, median, Q3, maximum, range, interquartile range

Initially, the dataset is loaded using the pandas library. It is loaded as follows:

imdbfile = pd.read\_csv("movie\_metadata.csv")



The PANDAS\_PROGILING gives the HTML profiling reports about the dataset

The DATASET INFO is as follows:

Number of varibles=28

Number of observations=5043

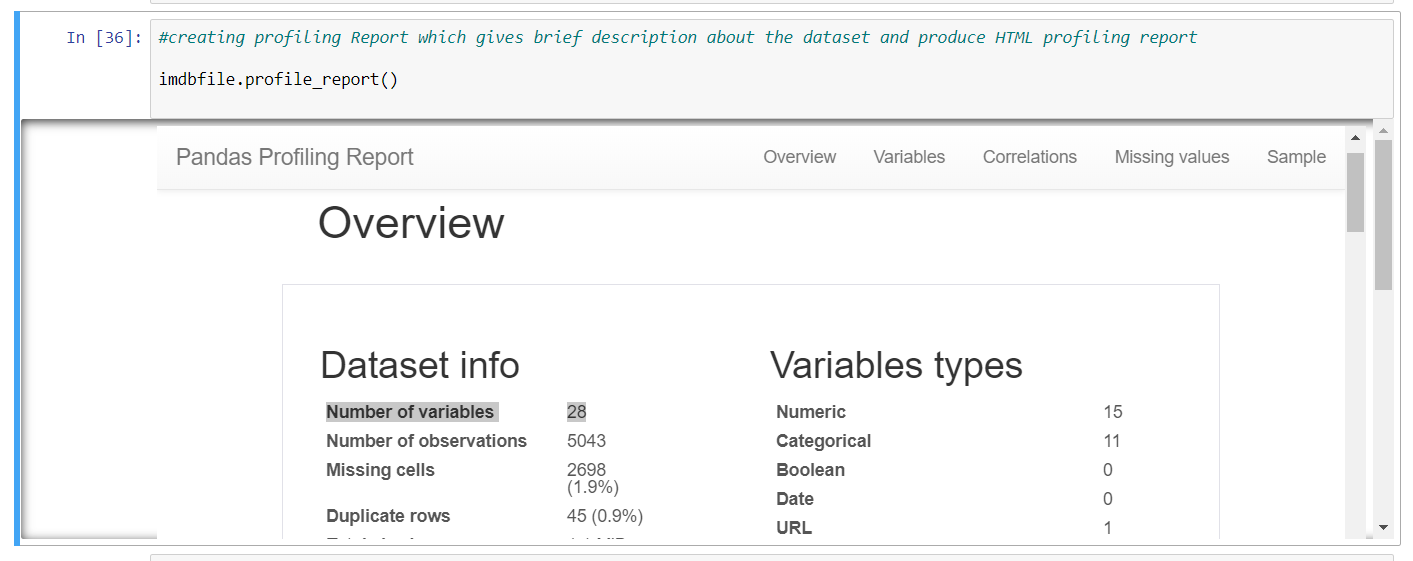
Missing values=2698

Duplicate Rows=45

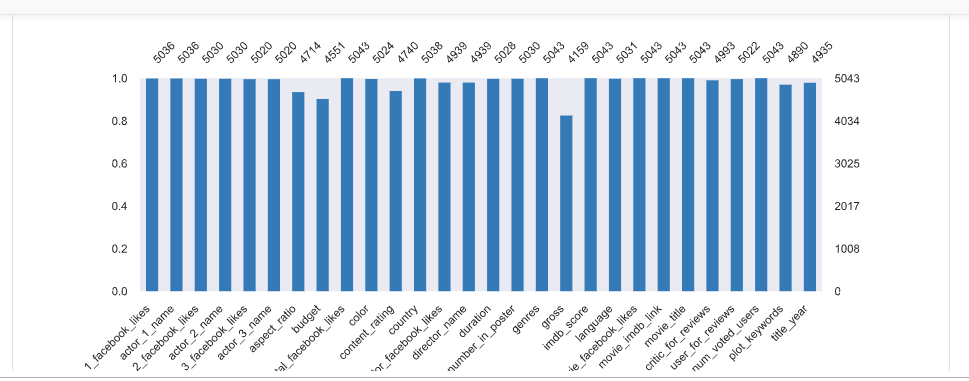
VARIABLE TYPES:

Numerical=15

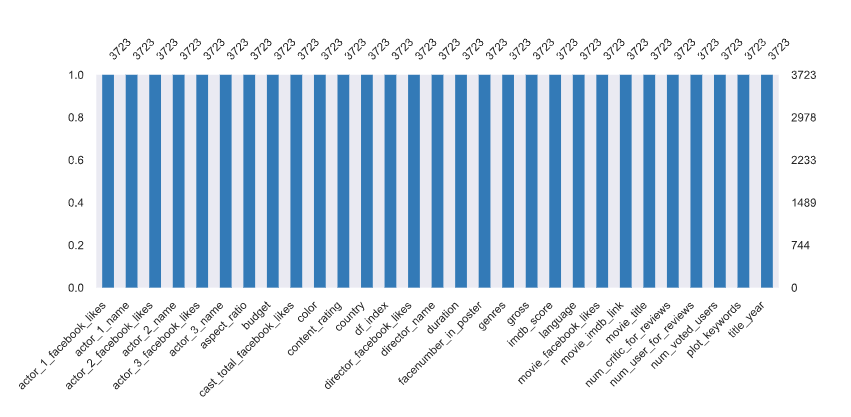
Categorical=11

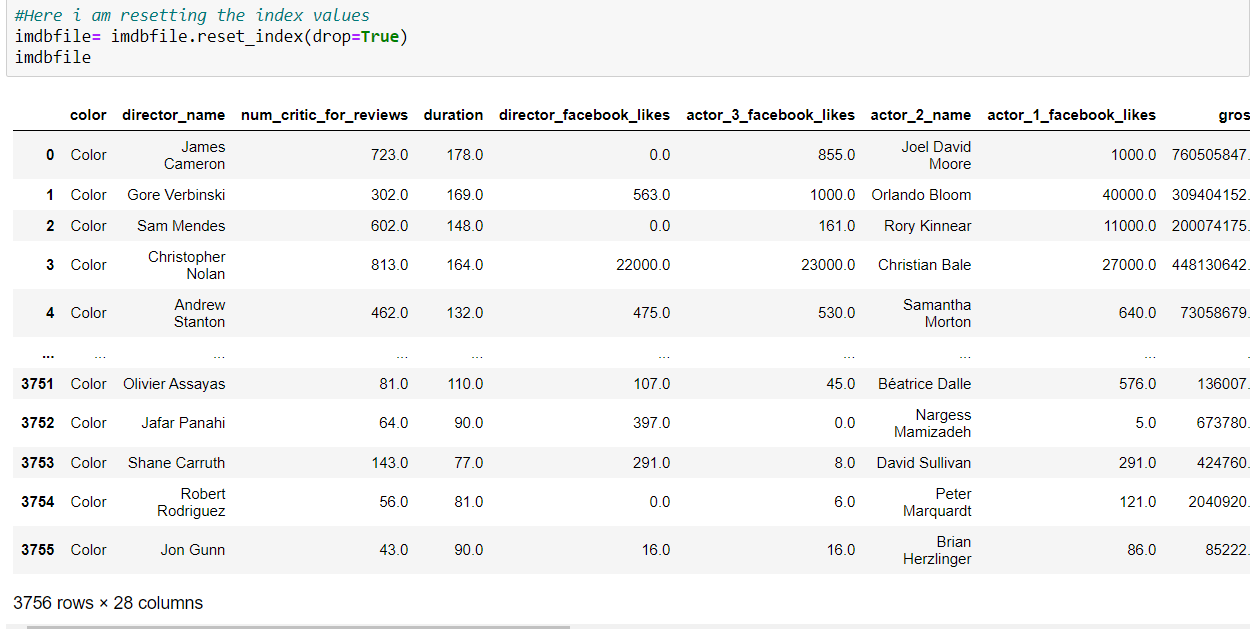


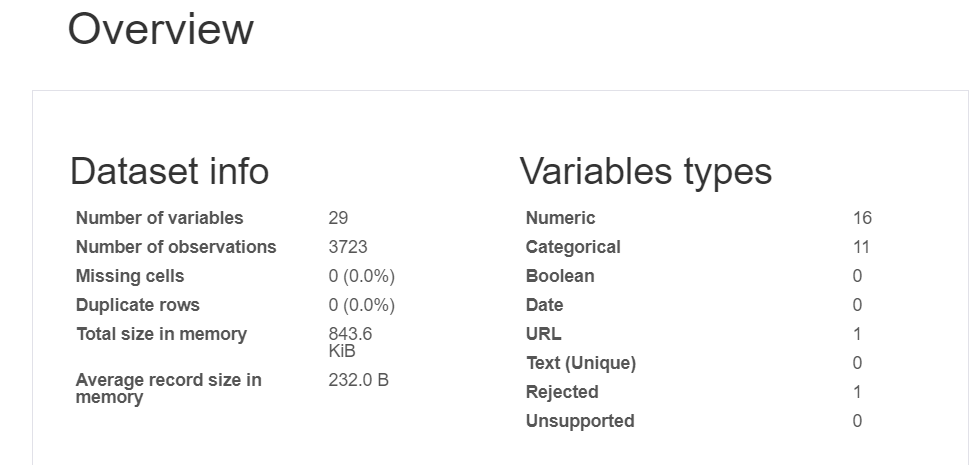
The missing data in Bar graph representation



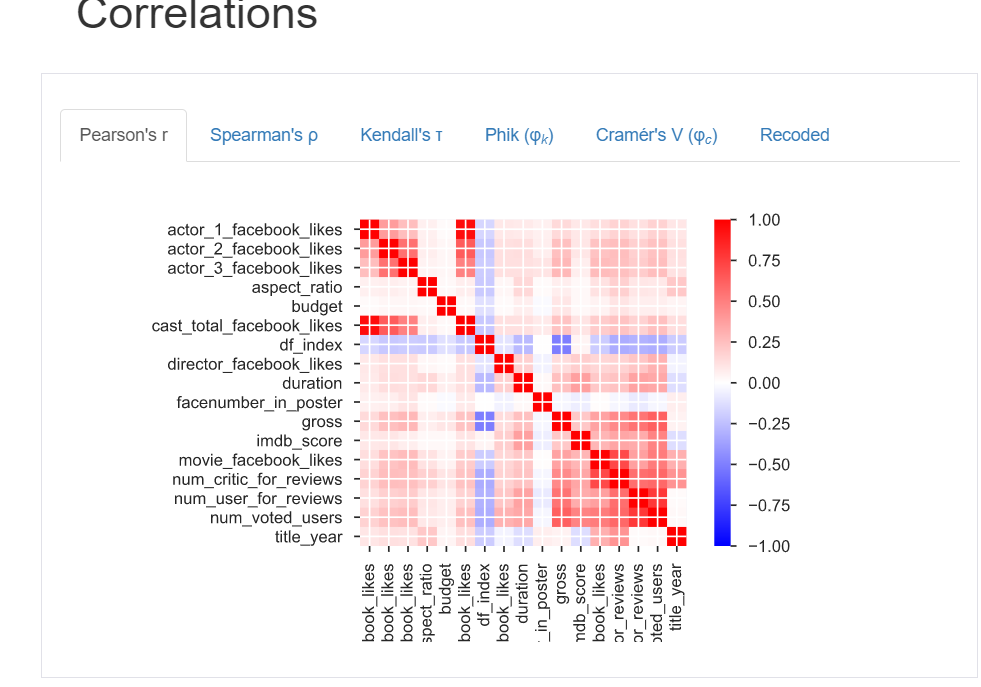
After loading the file I dropped all missing values the size of the dataset is (3756 rows × 28 columns) and I reset the index of the dataset.





And then deleted duplicate rows the shape of the dataset is (3723, 28)i.e 3723 rows and 28 columns.

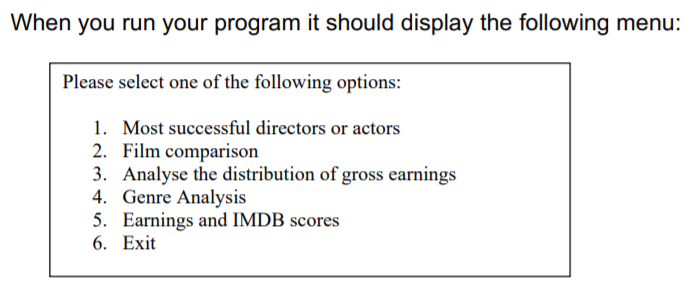
The correlations of the dataset are as follows after cleaning the given dataset.



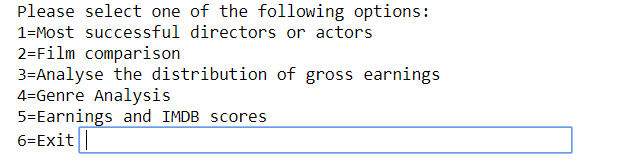
This is how I loaded the dataset and cleaned the dataset.

Now performing the given TASKS:

It is given that first, we need to display six options which are as shown in the figure



The main function should consist of these options when the main function is executed it should pop up this display, it is shown in the below figure



The first option is Most successful directors or actors when the user selects this option the main function will call the sub-function Most\_successful\_directors\_or\_actors(). Here two options will be displayed user should select one option and then enter the value of how many directors or actors he/she wants, the output will be displayed in the form of a horizontal bar graph.

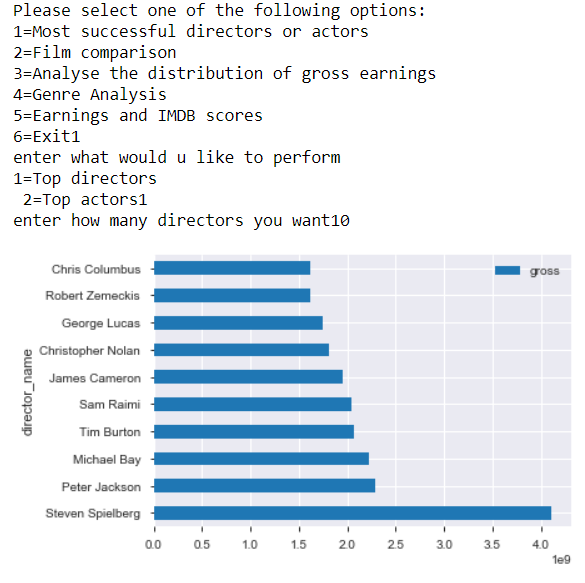
which will display the following options

1.Top directors

2.Top actors

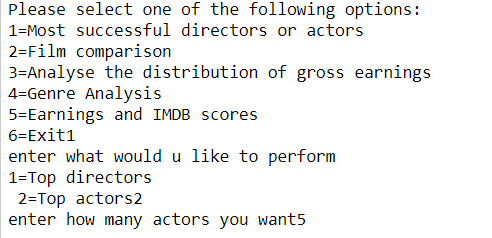
When the user selects the FIRST option it will ask how many top directors based on gross film earnings the user wants to see.

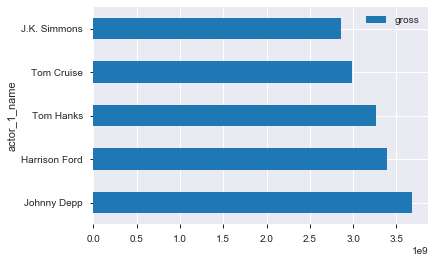
After Entering the value one horizontal bar graph will be displayed, which will give the top directors based on the gross film earnings.

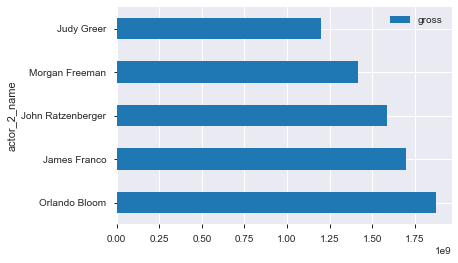


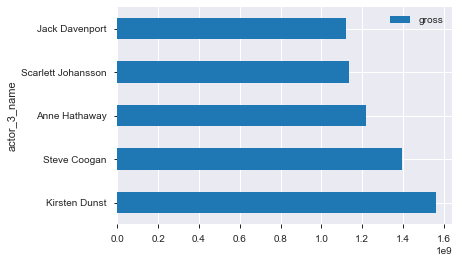
If the user selects the second option it will ask how many top actors based on gross film earnings the user wants to see.

After Entering the value one horizontal bar graph will be displayed, which will give the top actors based on the gross film earnings.Here there are three actor names columns I.e actor\_1\_name,actor\_2\_name,actor\_3\_name so three horizontal graphs appear one for each column.

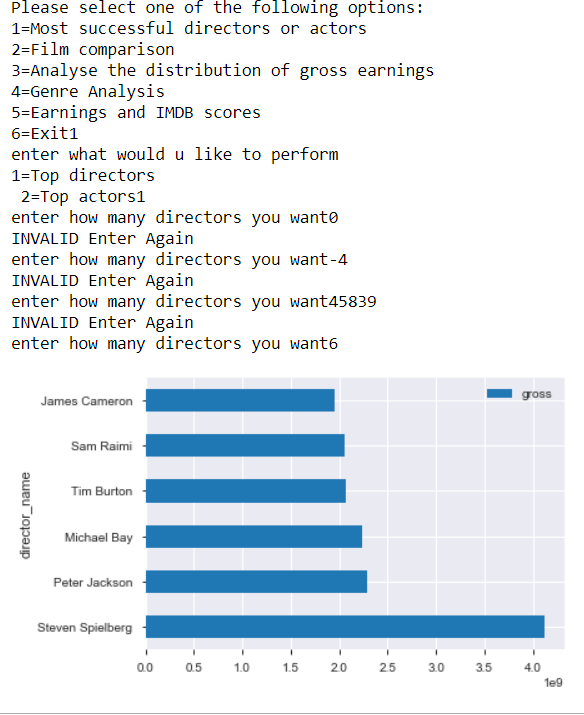








If the user enters the values out of the range then it will pop invalid enter again



This is how the first option is implemented.

If the user selects the SECOND option in the menu i.e Film comparison the main function will call the sub-function Film\_comparison(). Here a user should enter two film names and then 3 options will be displayed after giving film names, the user should select any one of the options and then the bar graph with respect films and to that particular option will be displayed.

which will display the following options

Enter the first Film name:

(User have to enter first film name)

Enter the second film name:

(User have to enter second film name)

1. IMDB Scores

2. Gross Earning

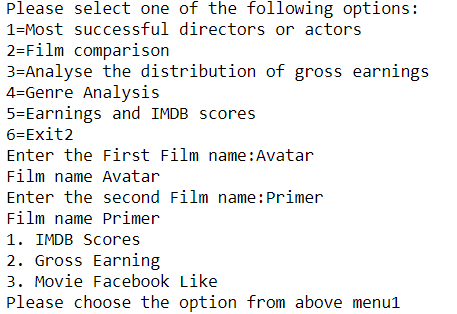
3. Movie Facebook Like

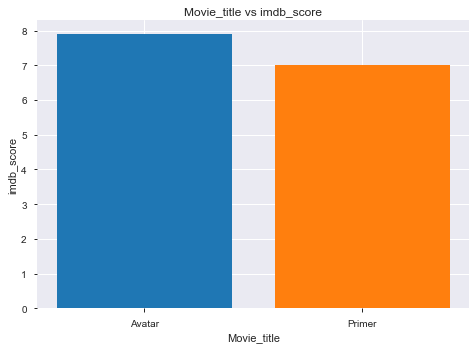
Please choose the option from the above menu

(User should choose one option among the three options)

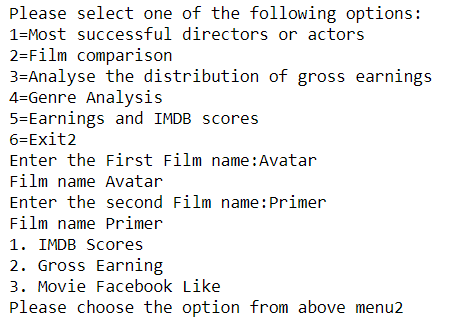
The Execution is as follows

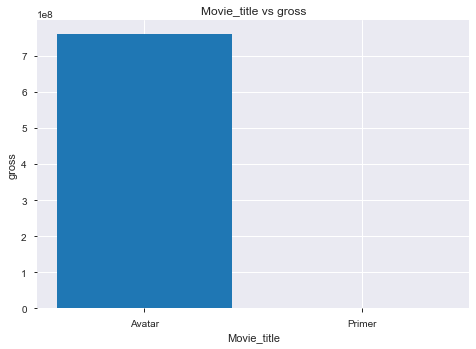
If the user selects the 1st option i.e IMDB score the output is as follows



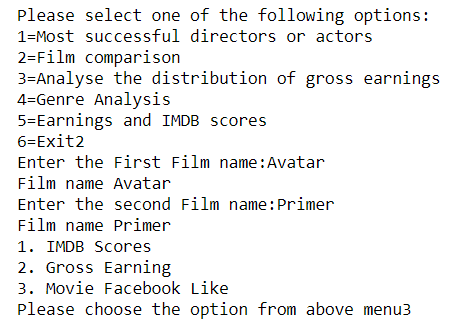


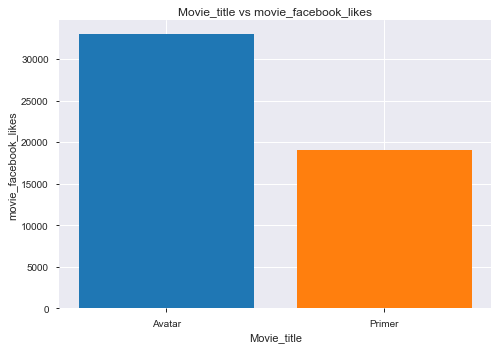
If the user selects the 2nd option i.e Gross Earning the output is as follows



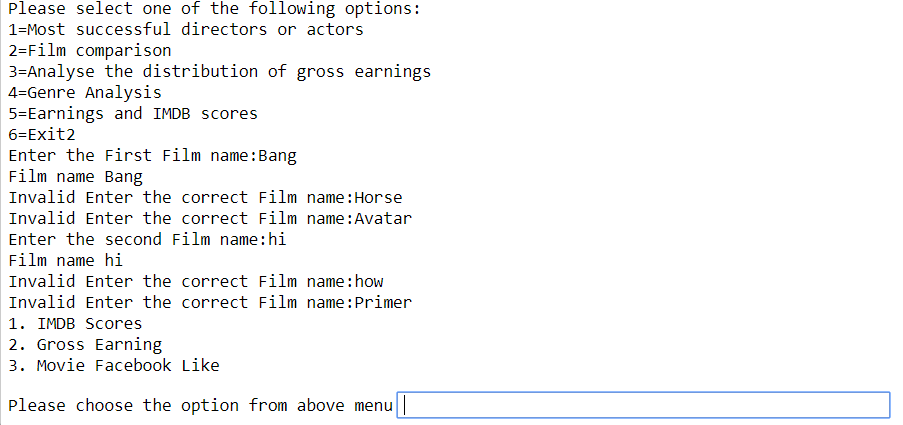


If the user selects the 3rd option i.e Movie Facebook Like the output is as follows





If the user selects the film which is not there in the dataset, it will ask the user that he entered films is invalid the correct film name



This is how the second option is implemented.

If the user selects the THIRD option in the menu i.e Analyse the distribution of gross earnings main function will call the sub-function Analyse\_the\_distribution\_of\_gross\_earnings (), which will display the following options

Enter the start year:

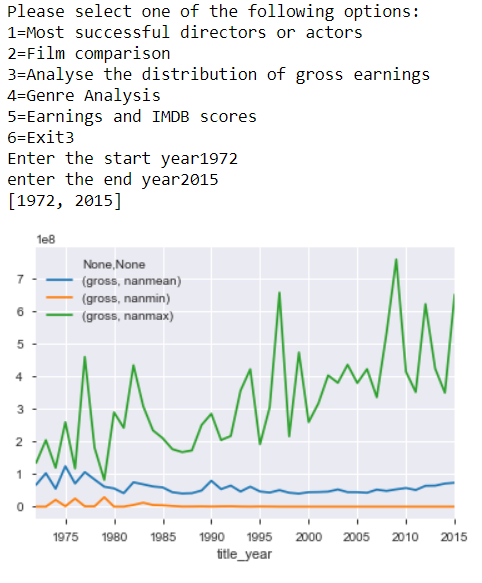
(User should enter the start year)

Enter the end year:

(User should enter the End year)

After that the start and end year will be displayed.

And then the minimum, maximum and average of gross film earnings will be displayed for all the years between the start and end year in the form of a line graph.



If the user enters the years which are not present it will display no data to display.

This is how the third option is implemented.

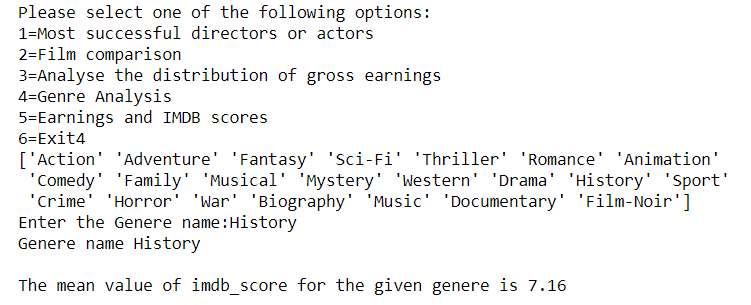
If the user selects the FOURTH option in the menu i.e Genre Analysis the main function will call the sub-function Genre\_Analysis (). Here a user should enter one Genere, and then the mean value of imdb\_score for the given genre will be displayed.

Firstly

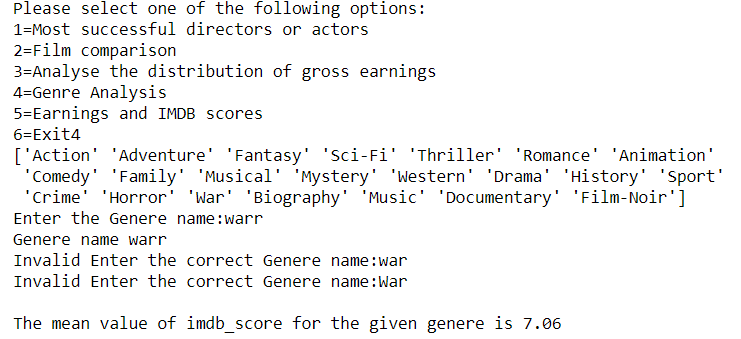
Enter the Genere name:

(User should enter the genre name)

Then the mean value will be displayed as follows



If the user enters the genre which is not present in the dataset, it will ask the user it is Invalid to enter the correct genre.



This is how the Fourth option is executed.

If the user selects the FIFTH option in the menu i.e Earnings and IMDB scores the main function will call the sub-function Earnings\_and\_IMDB\_scores ().

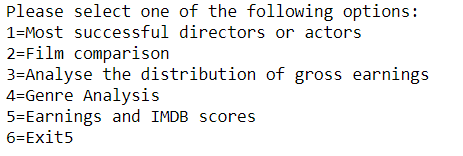
When this option is selected automatically all the graphs will display.

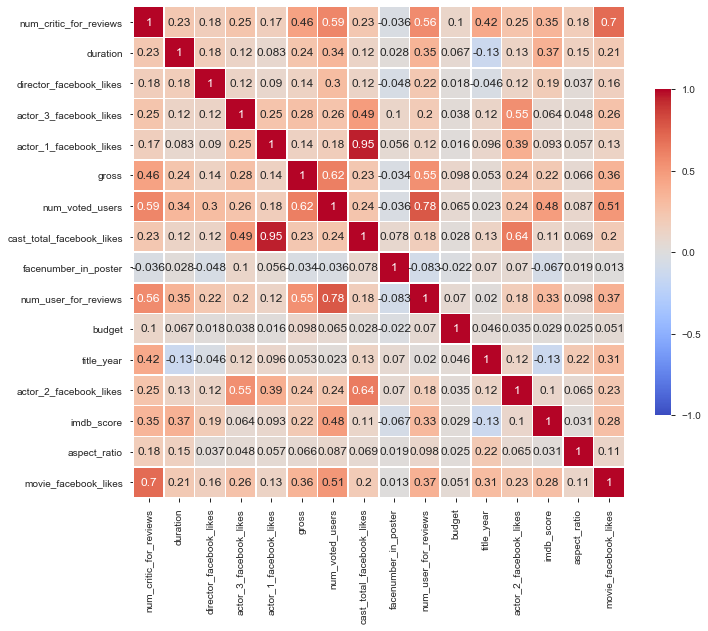
The first graph is the heatmap correlation graph, it will tell the relation between two variables.

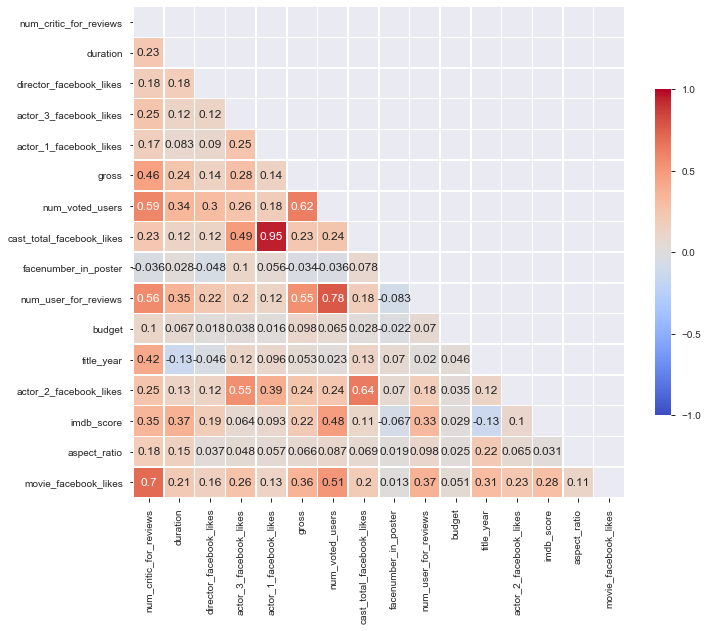
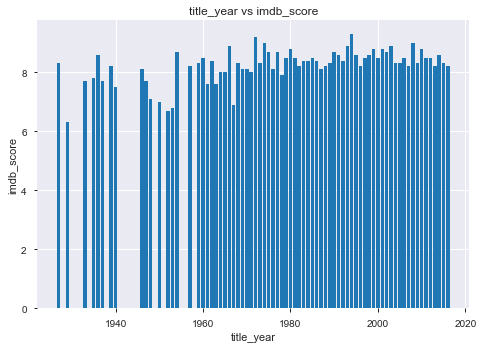
The second graph is also the heatmap correlation graph without the upper triangular matrix

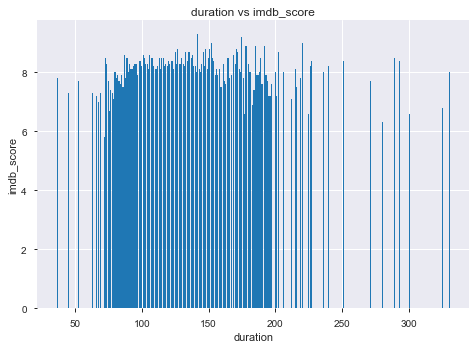
After this all the bar graphs with imdb\_score on the y-axis and the numerical variables on x\_axis.

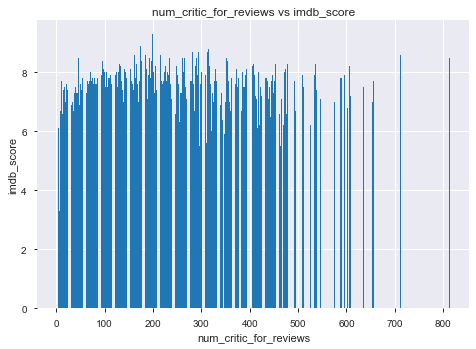
All the bargraphs between numerical IMDB scores and other numerical columns in the dataset are plotted.

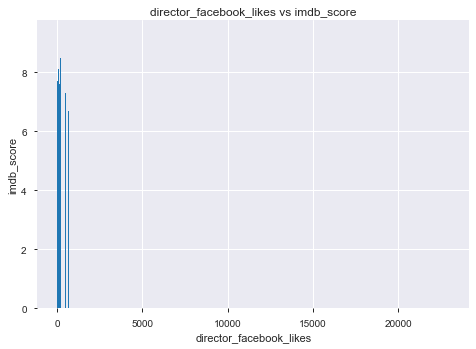


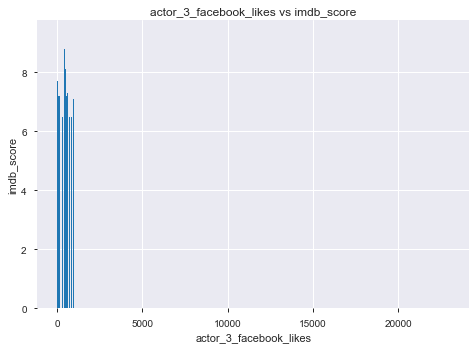


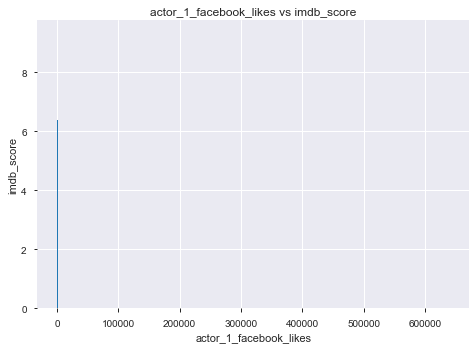
 

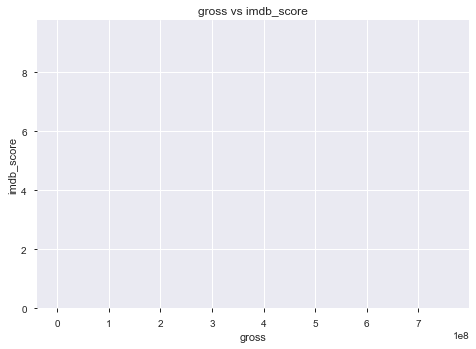


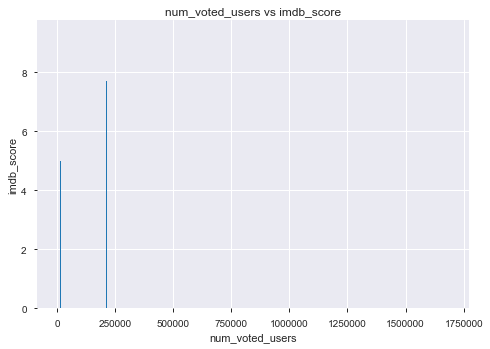


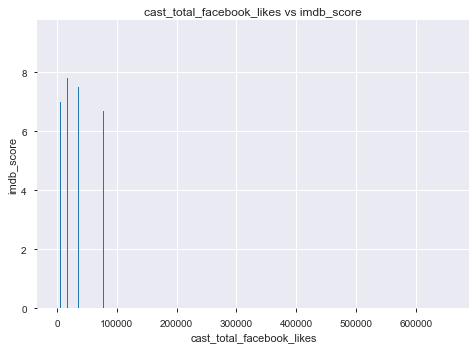


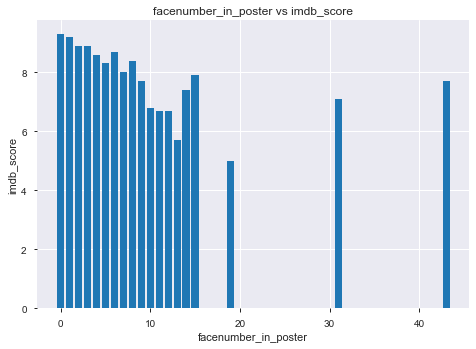


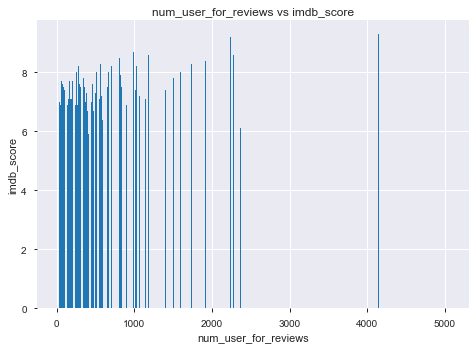


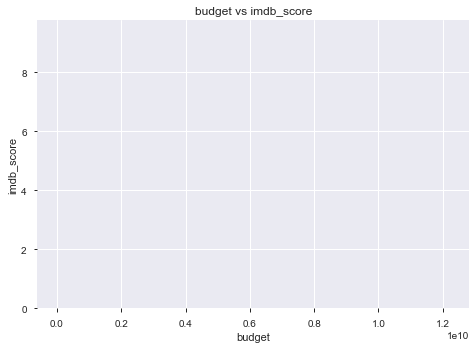


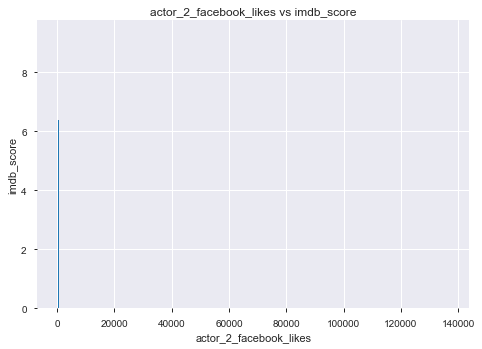


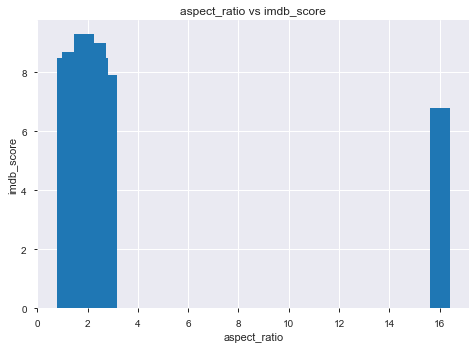


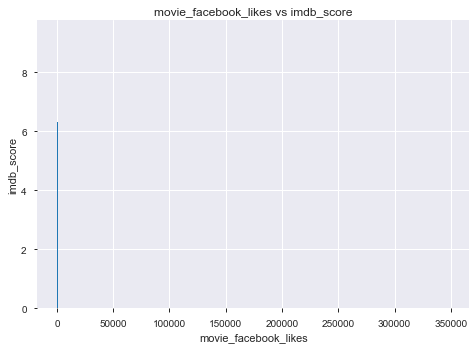












The variables

num\_voted\_users(0.48)

num\_user\_for\_reviews(0.38),duration(0.37),num\_critic\_for\_reviews(0.35)

movie\_facebook\_likes(0.28),gross(0.22)

director\_facebook\_likes(0.19),cast\_total\_facebook\_likes(0.11), actor\_2\_facebook\_likes(0.1), actor\_1\_facebook\_likes(0.093),aspect\_ratio(0.031),budget(0.029),actor\_3\_facebook\_likes(0.06),facenum\_in\_posts(-0.067), title\_year(-0.13)

The num\_voted\_users, num\_user\_for\_reviews, duration, num\_critic\_for\_reviews, movie\_facebook\_likes, gross, director\_facebook\_likes, cast\_total\_facebook\_likes, actor\_2\_facebook\_likes,actor\_1\_facebook\_likes,aspect\_ratio,budget,actor\_3\_facebook\_likes is having positive correlation with imdb\_score .

Whereas facenum\_in\_posts, title\_year have a negative correlation with imdb\_score.

In order to build a model, it is good to build with positive correlation variable with imdb\_score

Among all the positive correlation variables it is better to use num\_voted\_users, num\_user\_for\_reviews, duration, num\_critic\_for\_reviews, movie\_facebook\_likes, gross variable as these variables have correlation values greater than 0.1 and also comes under strong positive correlation.so it is better to build the model with these variables

This is how the Fifth option is implemented.

If the user selects the SIXTH option in the menu i.e EXIT, it will print that the user selected the EXIT option.

This is how all the SIX OPTIONS in the main function is executed.