Givaldo

R00182595

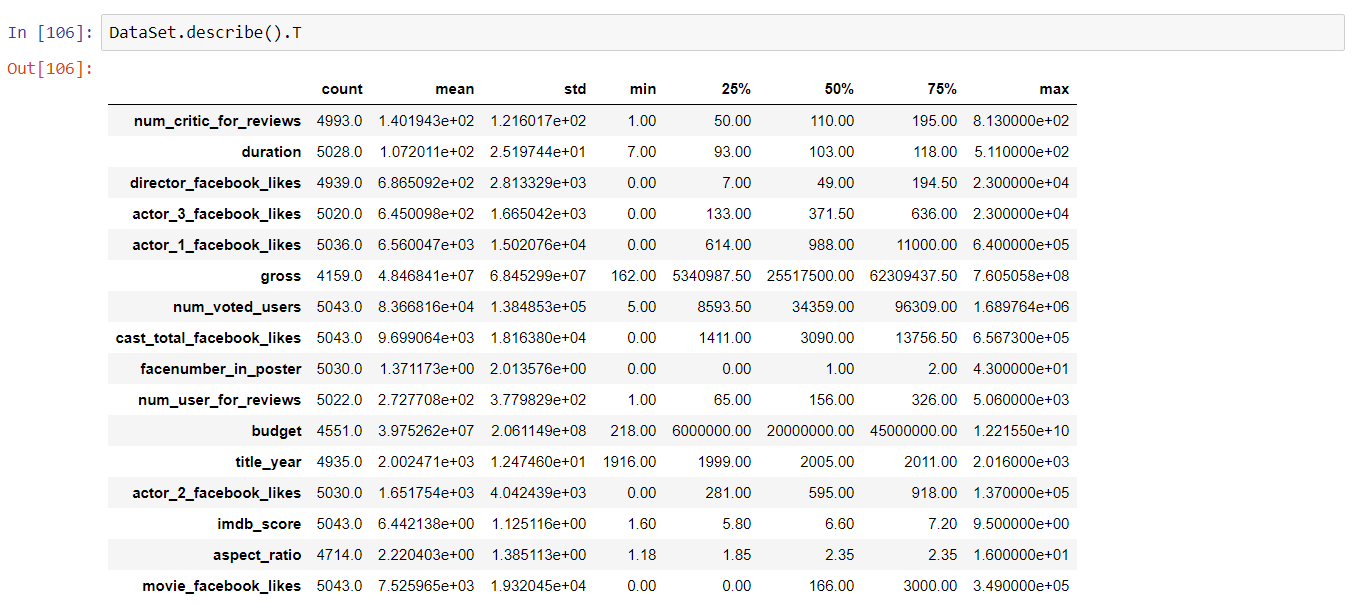
Guidelines and Submission Instructions:

1. The project is worth 40% of your overall module grade. You will produce a **jupyter notebook file** containing all your code and output for each of the operations outlined the project.=
2. Upload your solution jupyter file to Canvas before on or before 20
3. Go to the Assignment Project -I in Canvas to upload your file.
4. Once you have submitted your files you should verify that you have correctly uploaded them. It is your responsibility to make sure you upload the correct files.
5. Please make sure you **fully comment your code**. You should clearly be explaining the operation of important lines of code.
6. Please put your student name and number as comments at the top of your file.

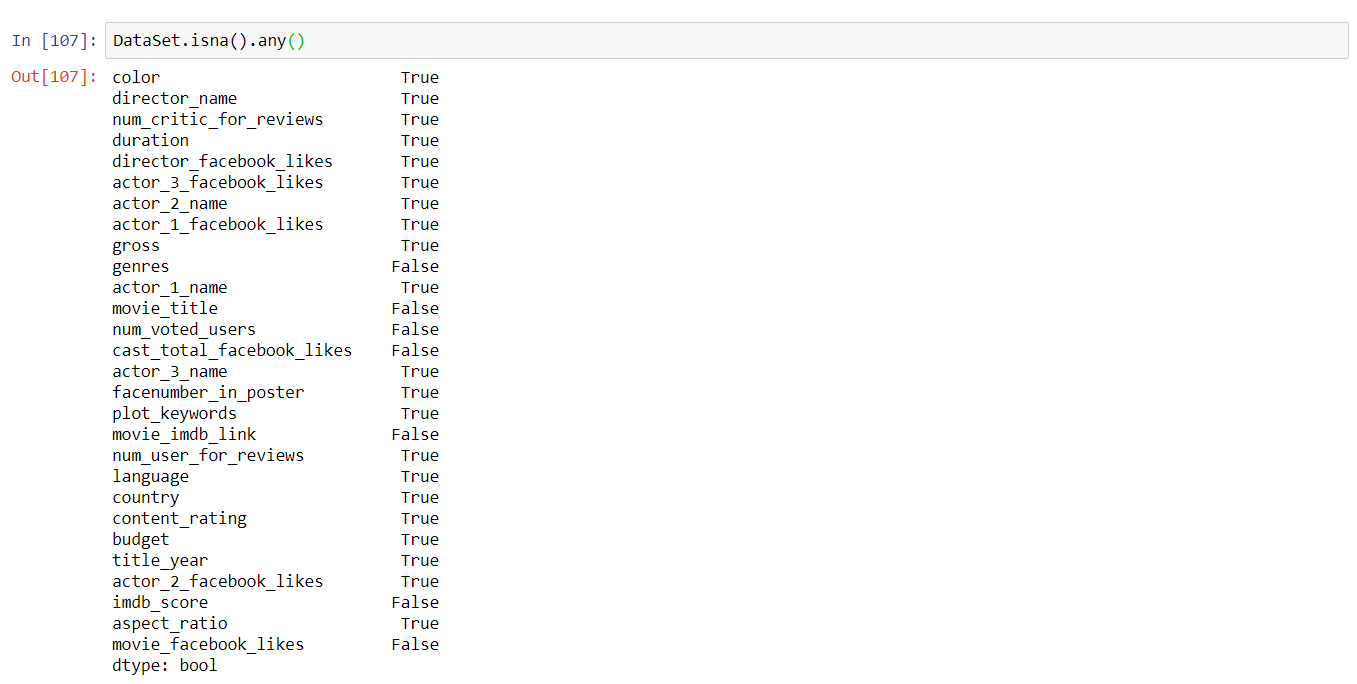
Import dataset and libraries

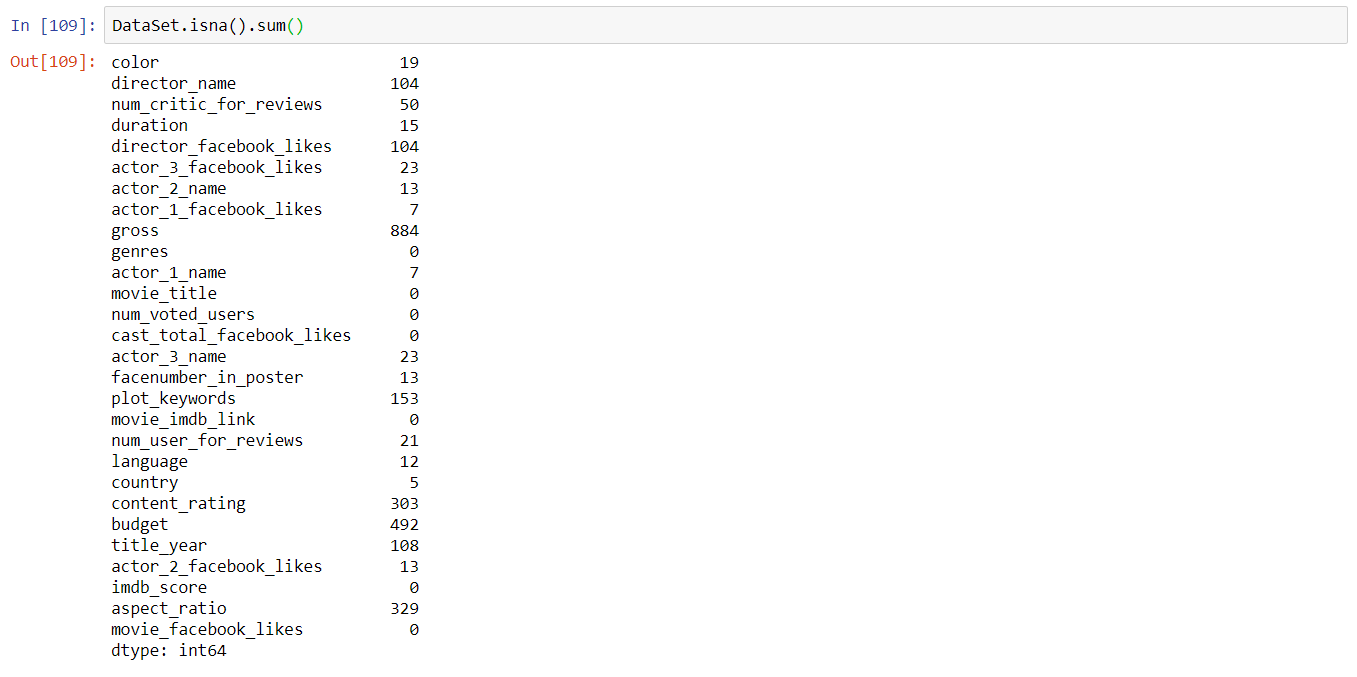


Describes() shows basics information about the dataset



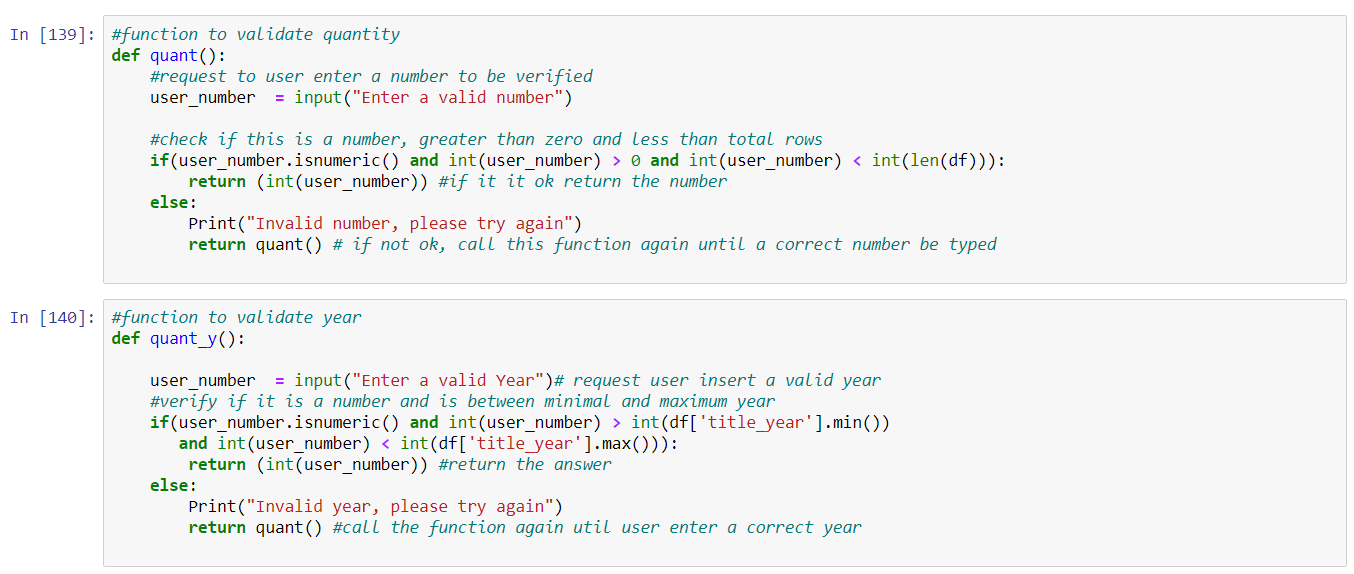
Checking information about missing values



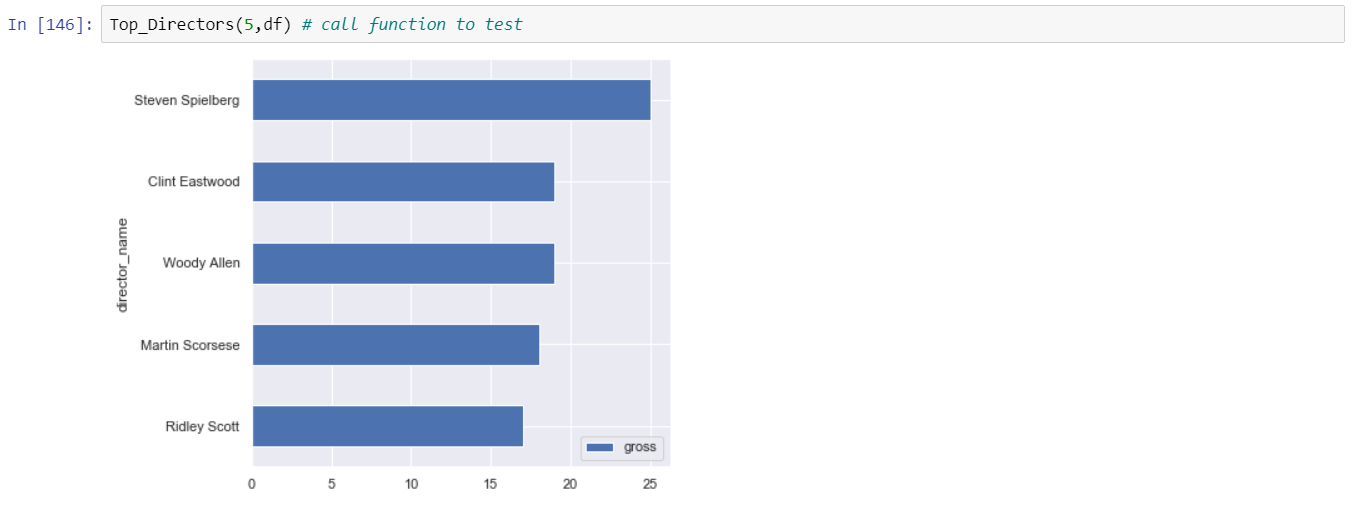


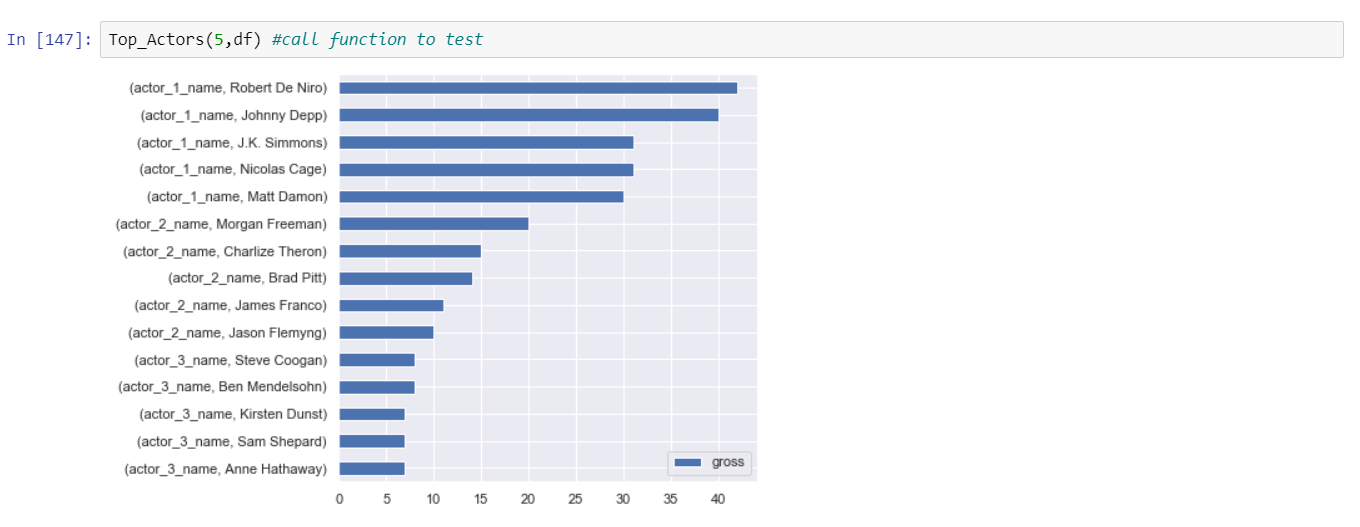
1. **Menu Option 1 – Most successful directors or actors**

The function quant() does some simple checks to verify if the number is correct



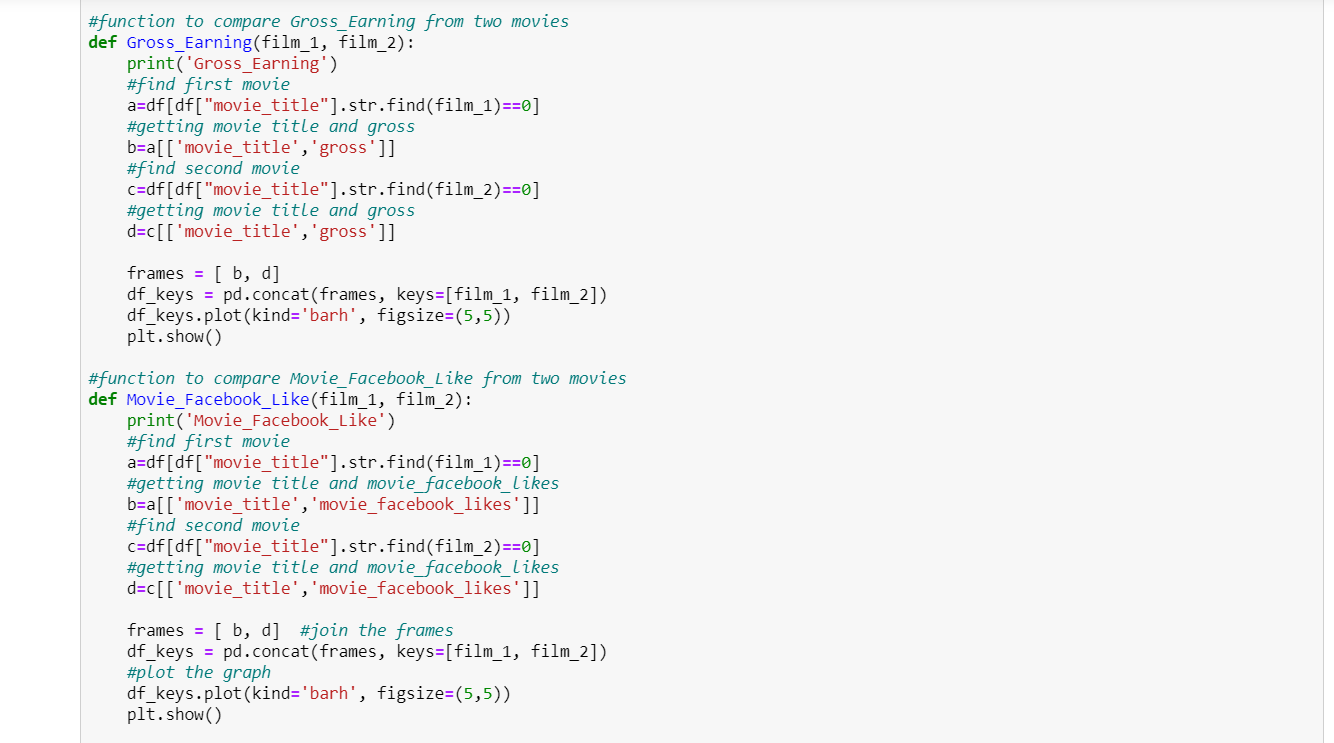
Both functions will return the total of Directors or actors according to what the user selects.

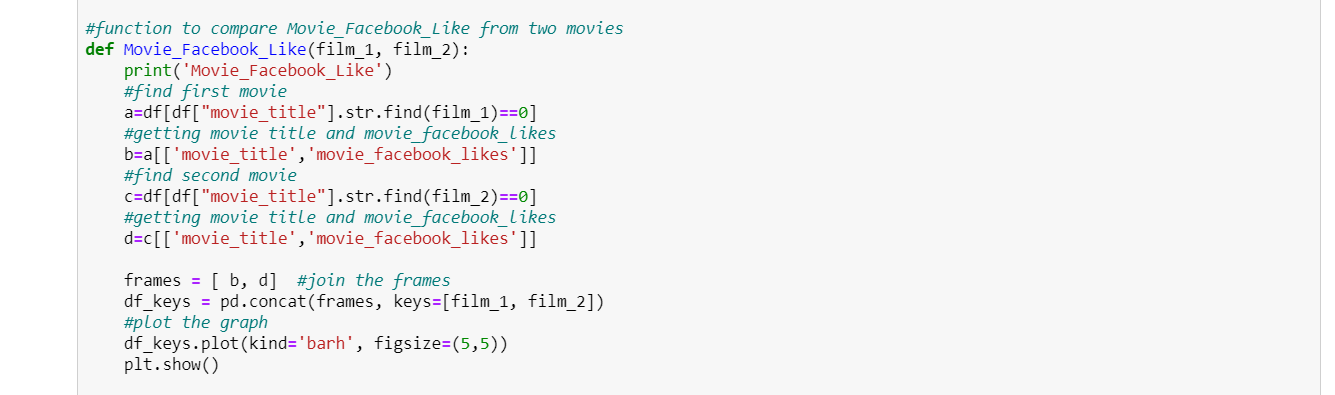
A simple test of the function.



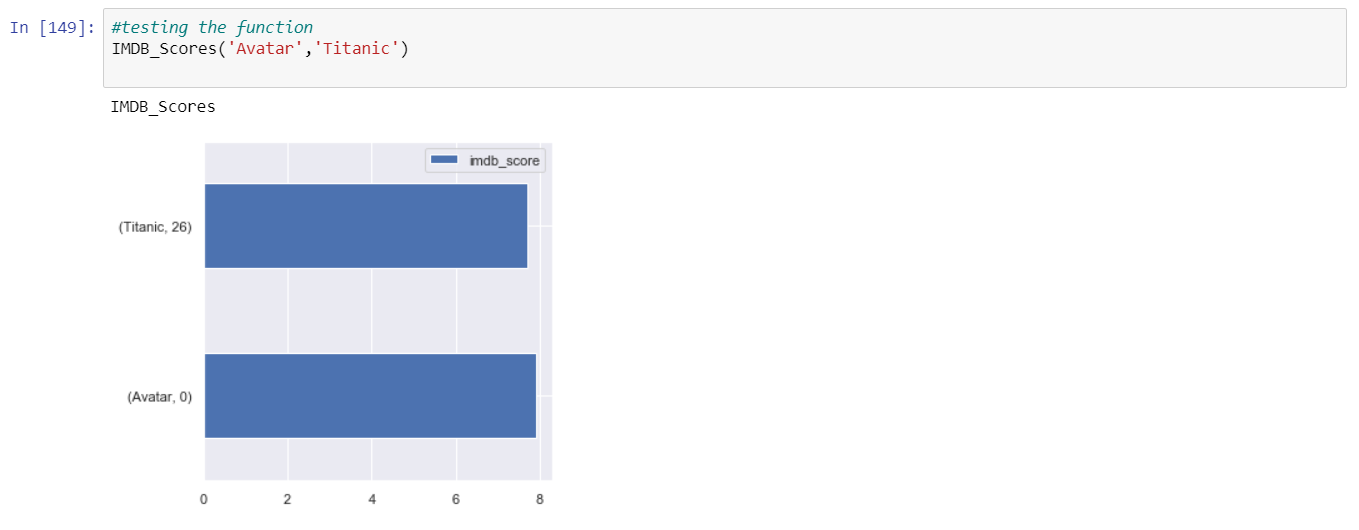
**2. Menu Option 2 – Film Comparison**

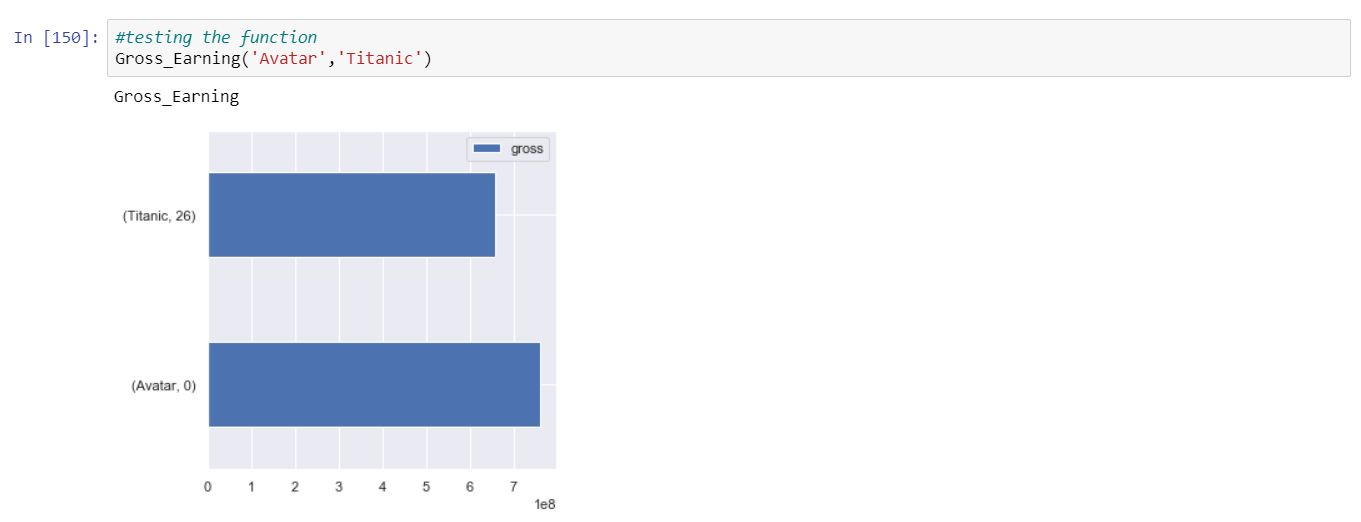
In these three functions, users will provide two film names and the function will return an IMDB scores comparison. There is a check in the movie name, to see if that name exists in the data set.

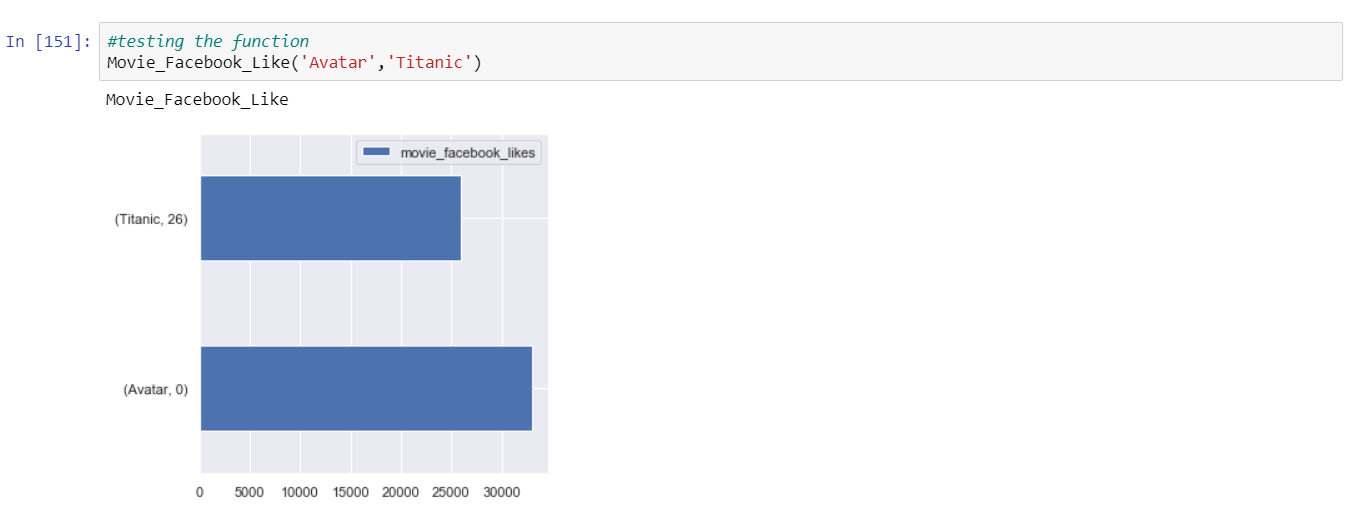




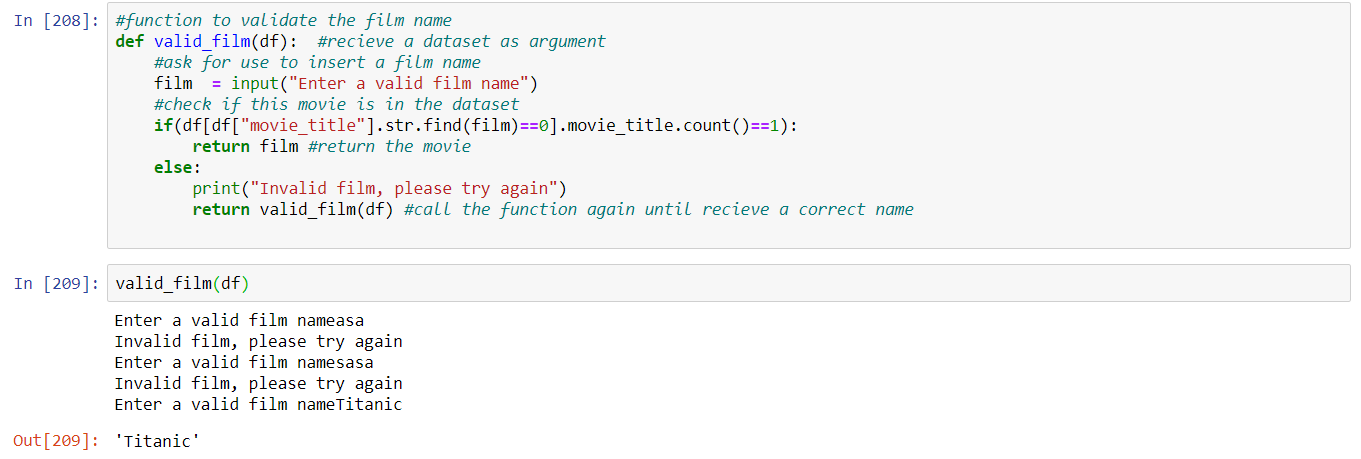
Test of the function.



As we can see, despite the fact both movies have almost the same IMDB score, the gross earnings are different. If we analyze, we can see that the gross earning has increasing with the time.

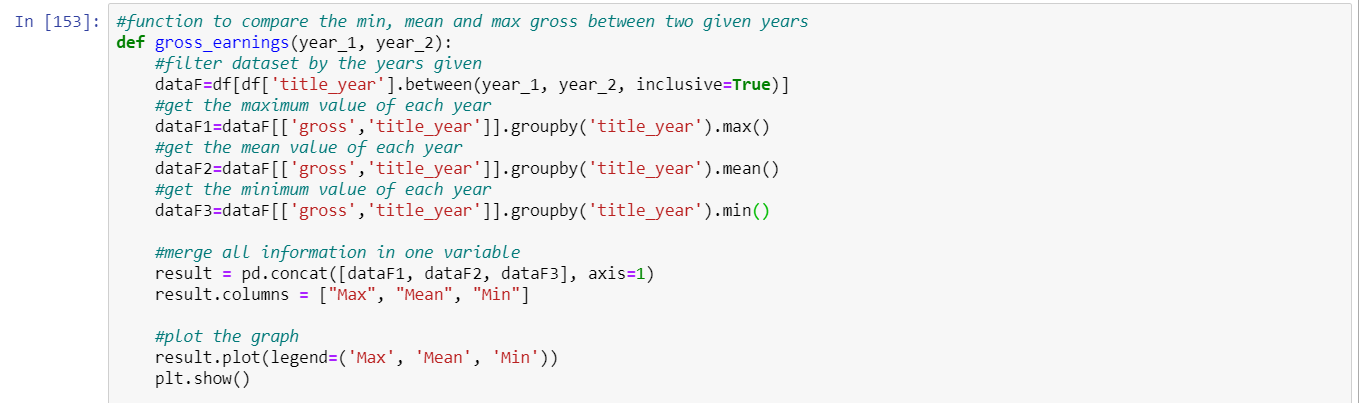


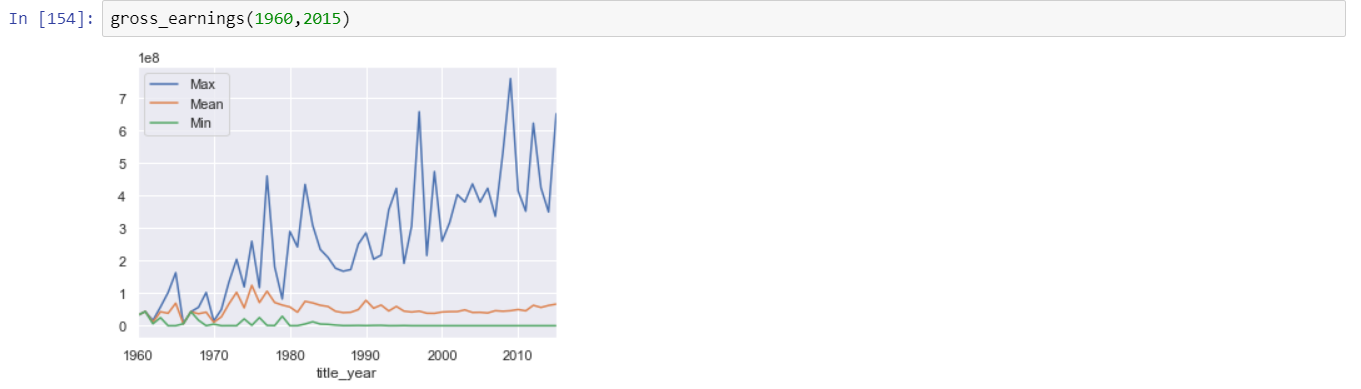
This function check if film imputed by ser exist in the data set.



**3. Menu Option 3 – Analyze the distribution of gross earnings**

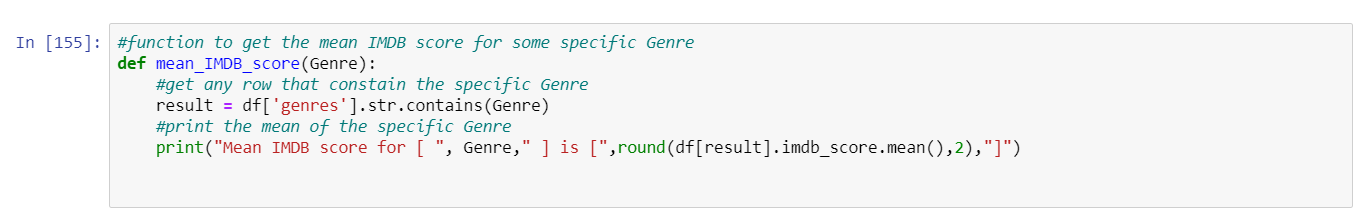
This function request two years and provide a line graph about max mean and min gross earning. We can see in the graph that the max has increased a lot, but the mean and min not.

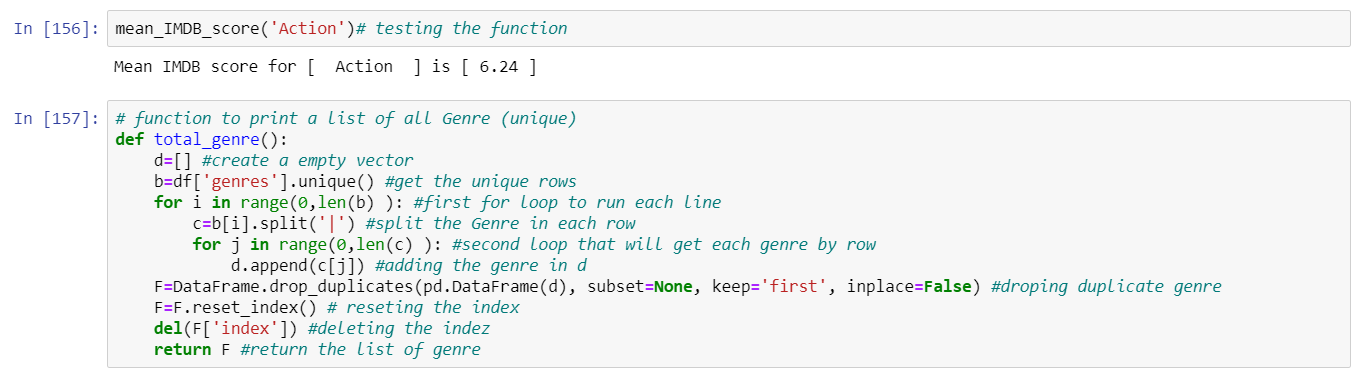




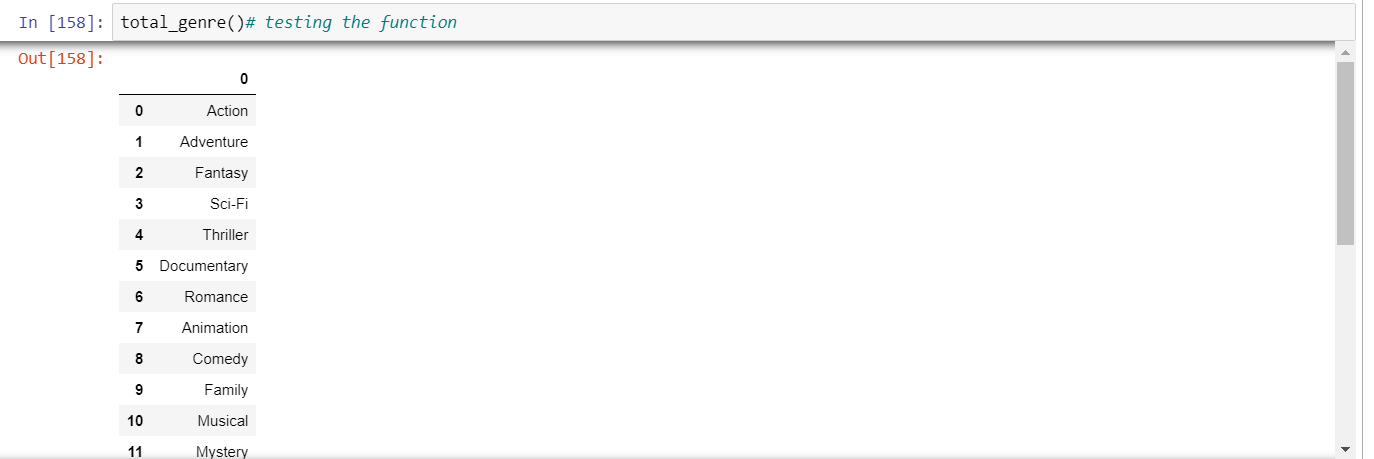
**4. Menu Option 4 – Genre Analysis**

This function will provide the mean of IMDB score for a genre inserted by the user.



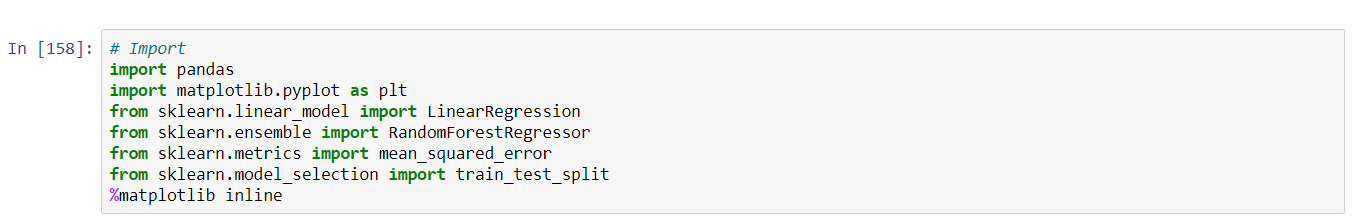


This function returns total (unique) genres

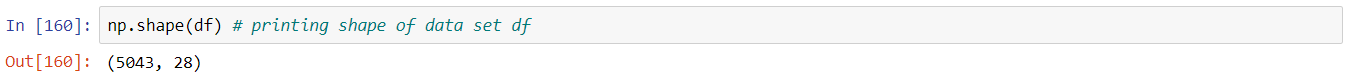


**5. Menu Option 5 – Earnings and IMDB scores**

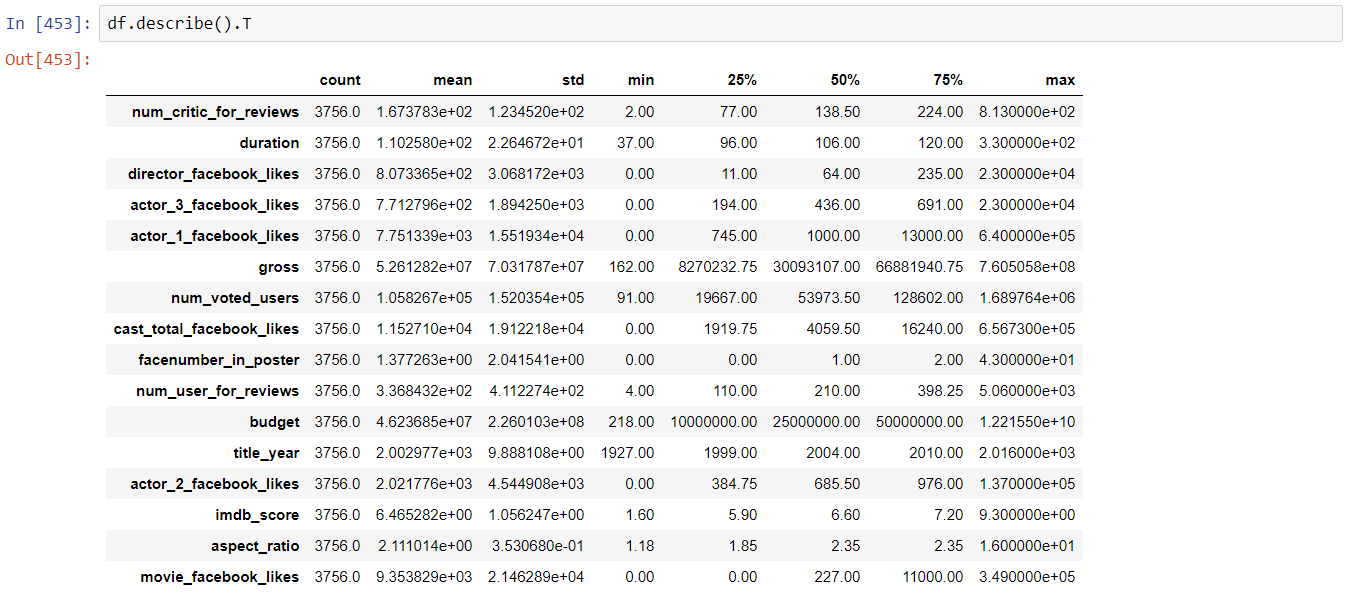
This is a previous analysis of this data set to find the best features to get a better accuracy for the model.

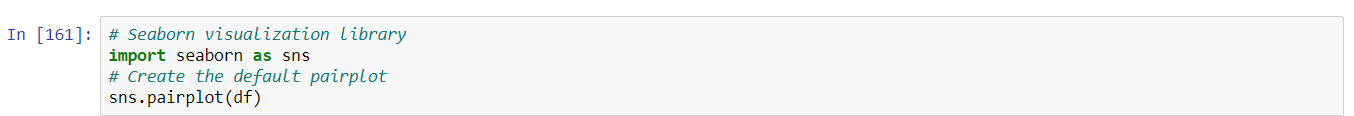


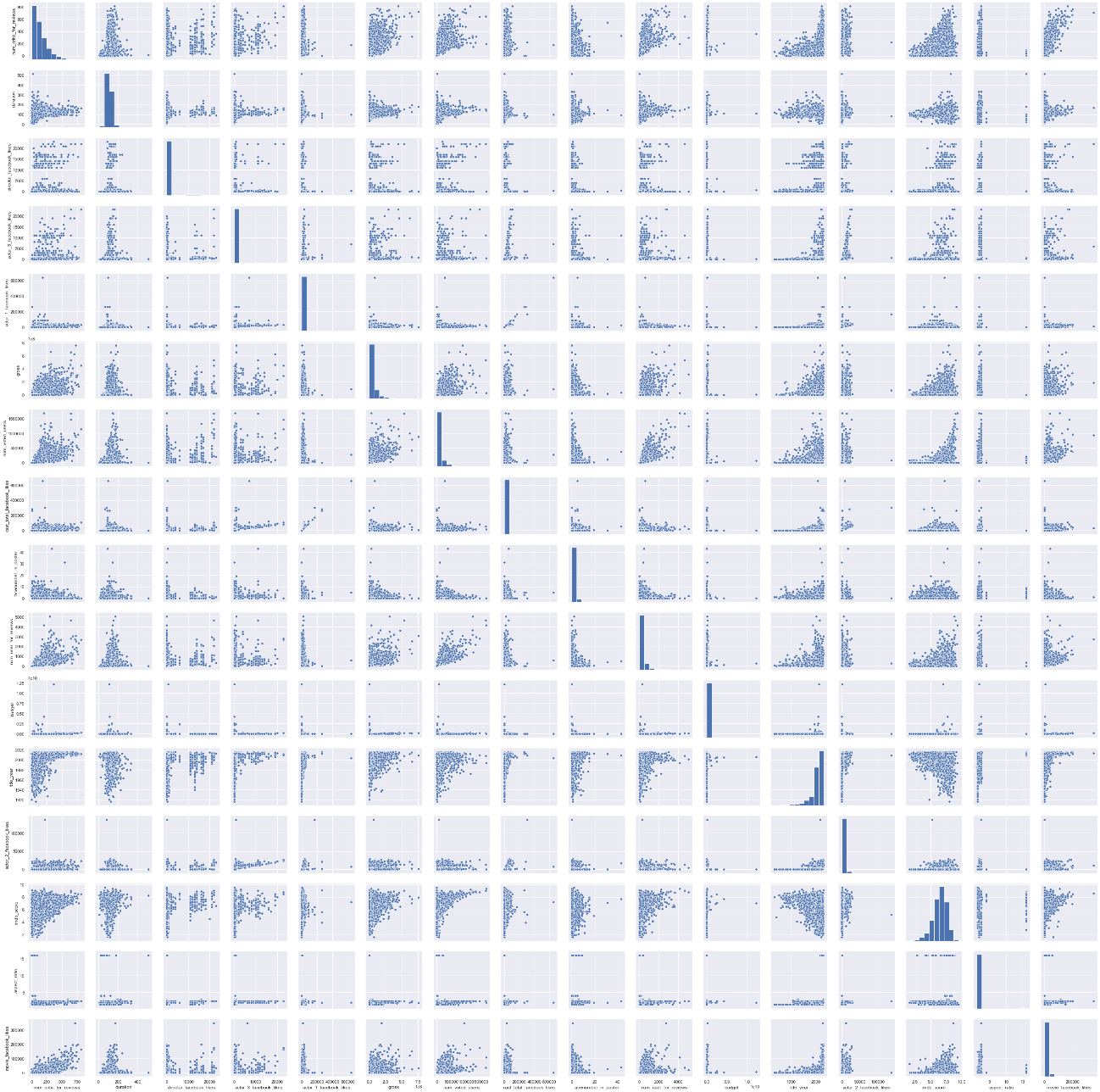




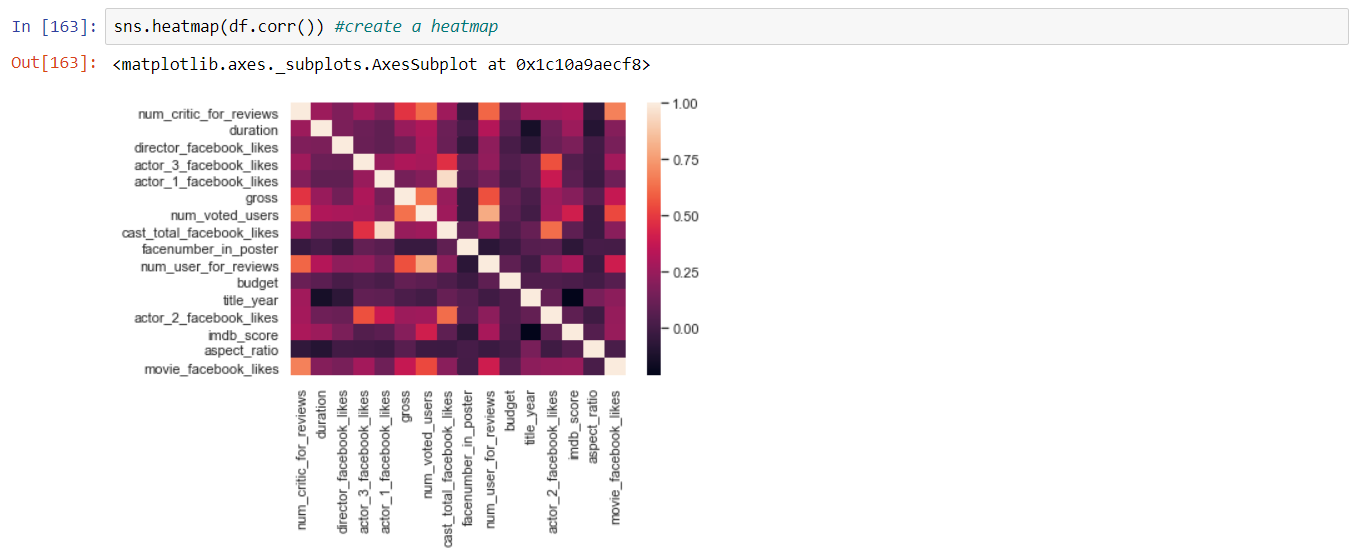
Describe function give me a overview of all variables.

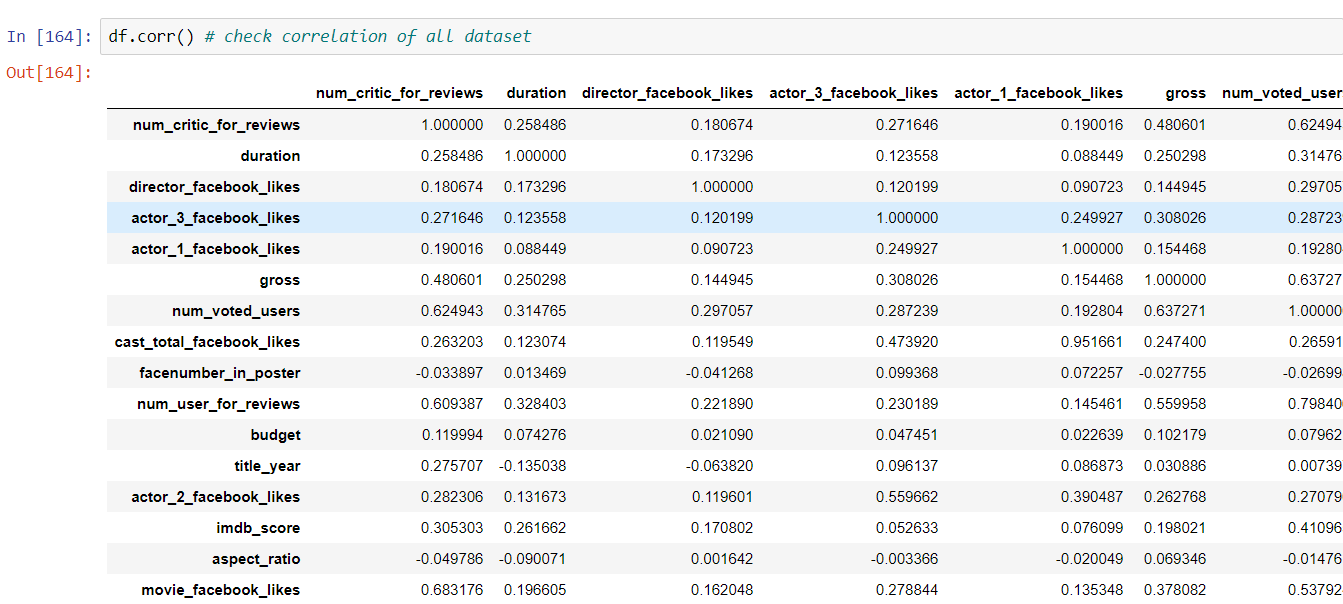


This plot gives us a scatterplot visualization of each variable with all others.



This head map is good to visualize the correlation between all variables, by the colors we can verify it faster than by just number.



Get the correlation by the number we can have more precise information.

We can get the correlation with a specific variable, in this case that is our target variable.

The highest correlation with imdb score

'num\_critic\_for\_reviews',

'duration',

'director\_facebook\_likes',

'gross',

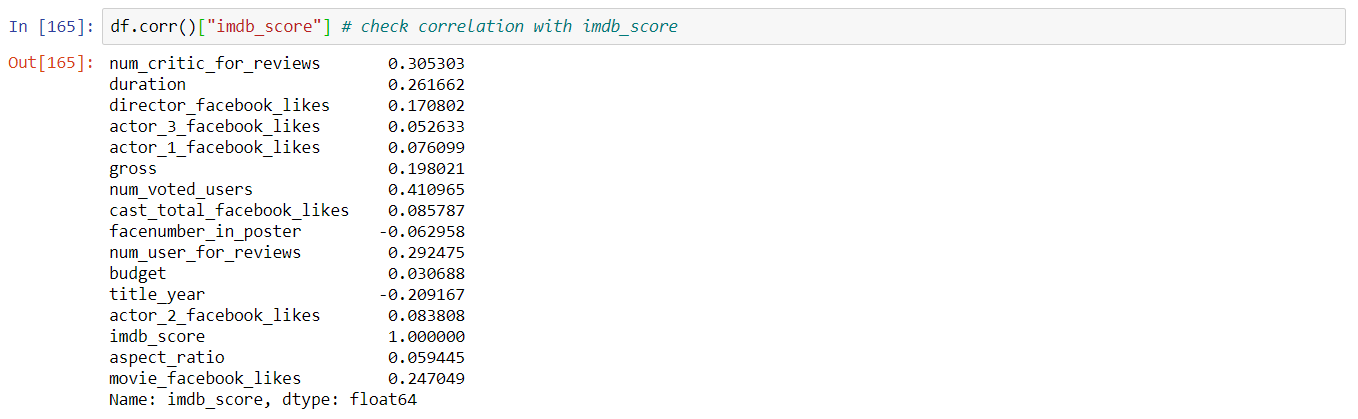
'num\_voted\_users',

'num\_user\_for\_reviews',

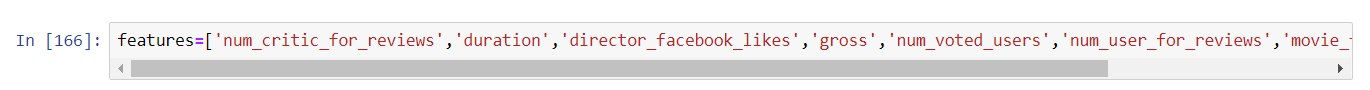
'movie\_facebook\_likes',

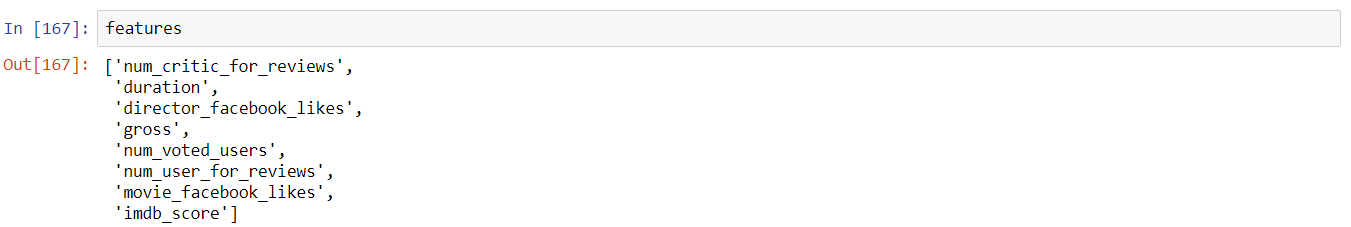
'cast\_total\_facebook\_likes',

'actor\_2\_facebook\_likes'



This features have the highest correlation with our target.

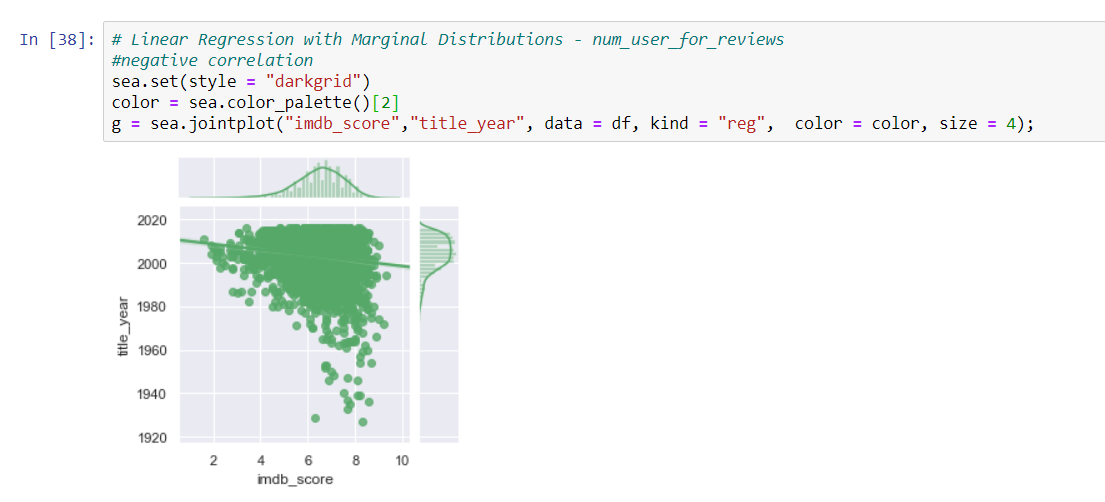




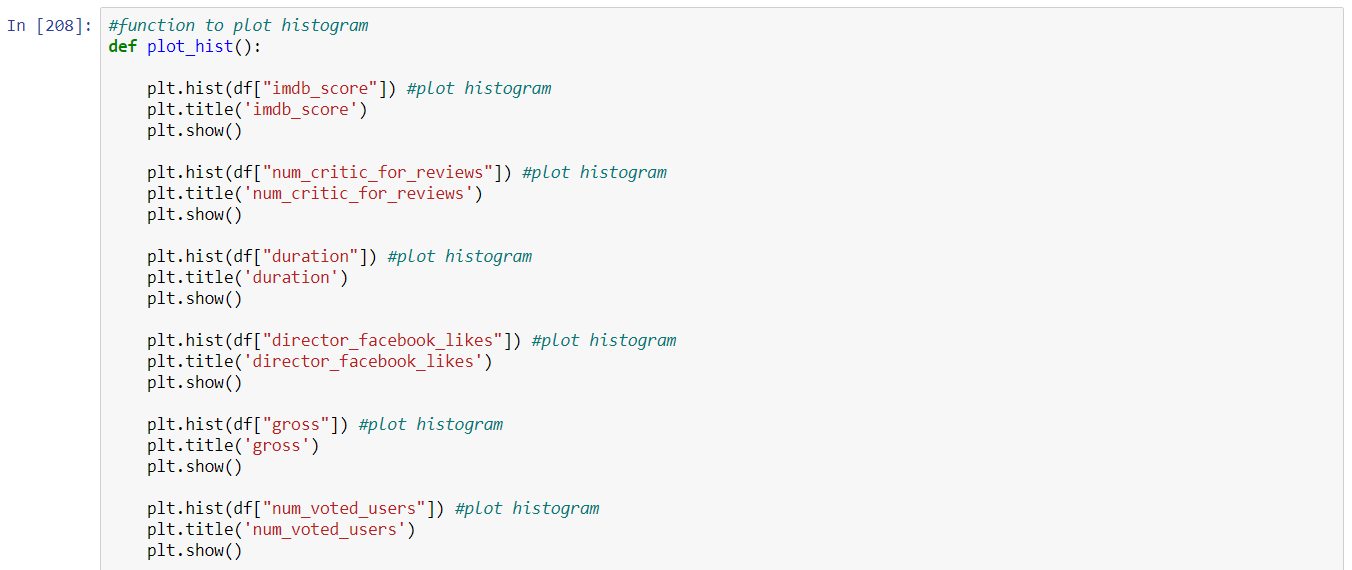
The scatterplot is good to see a correlation between two variables, in the examples below we can see a positive correlation with IMDB score, this means as much theses variables increase, imdb score will increase too.



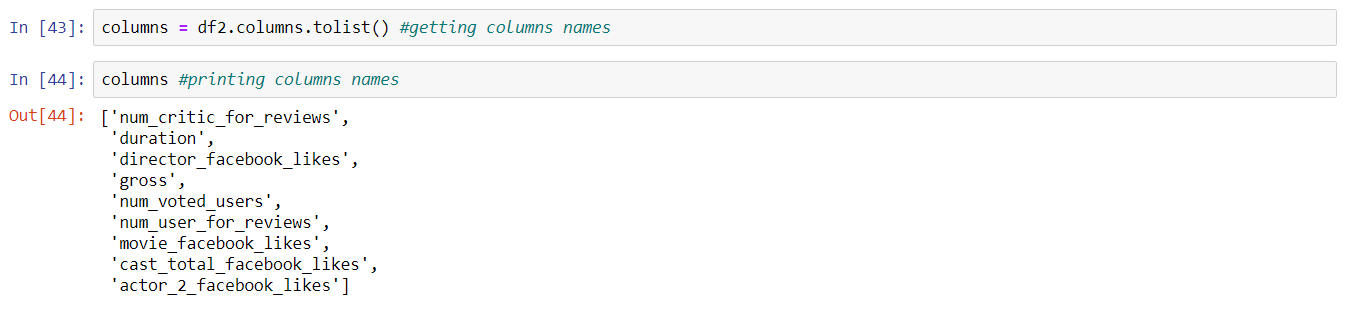
Example of negative correlation



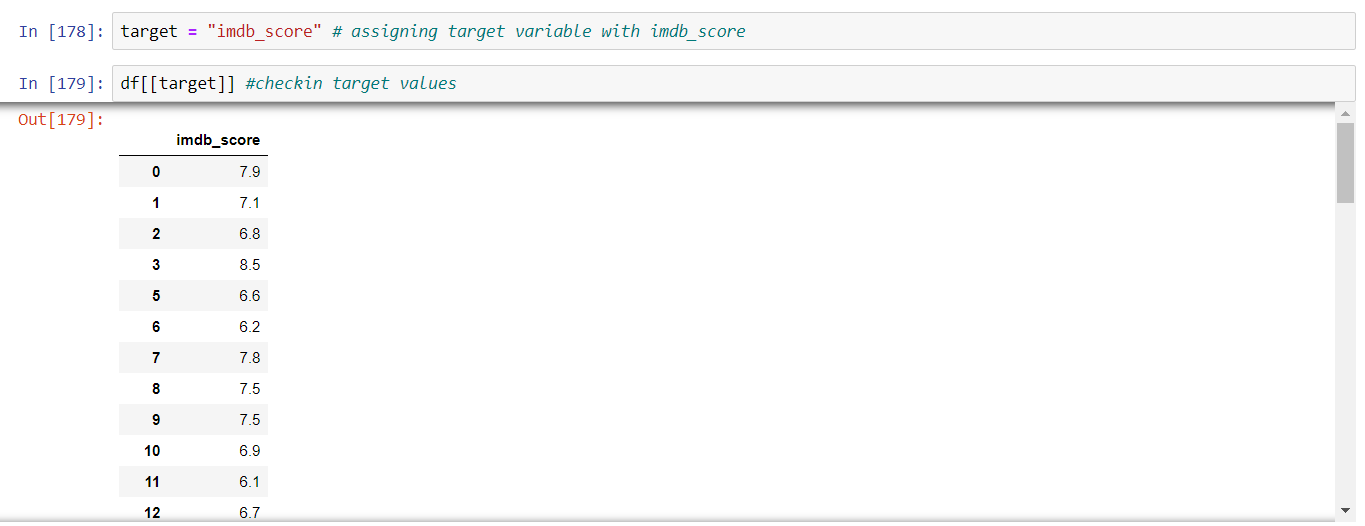
The histogram is good to visualize the distribution of some variable and where is concentrated most part of elements.





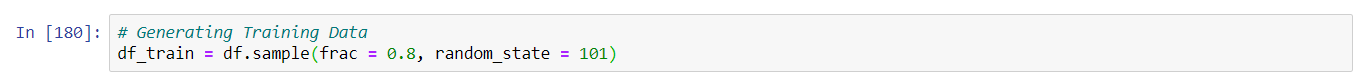


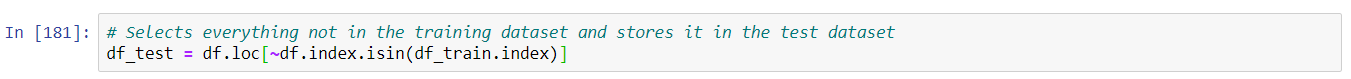
Our target is imdb score



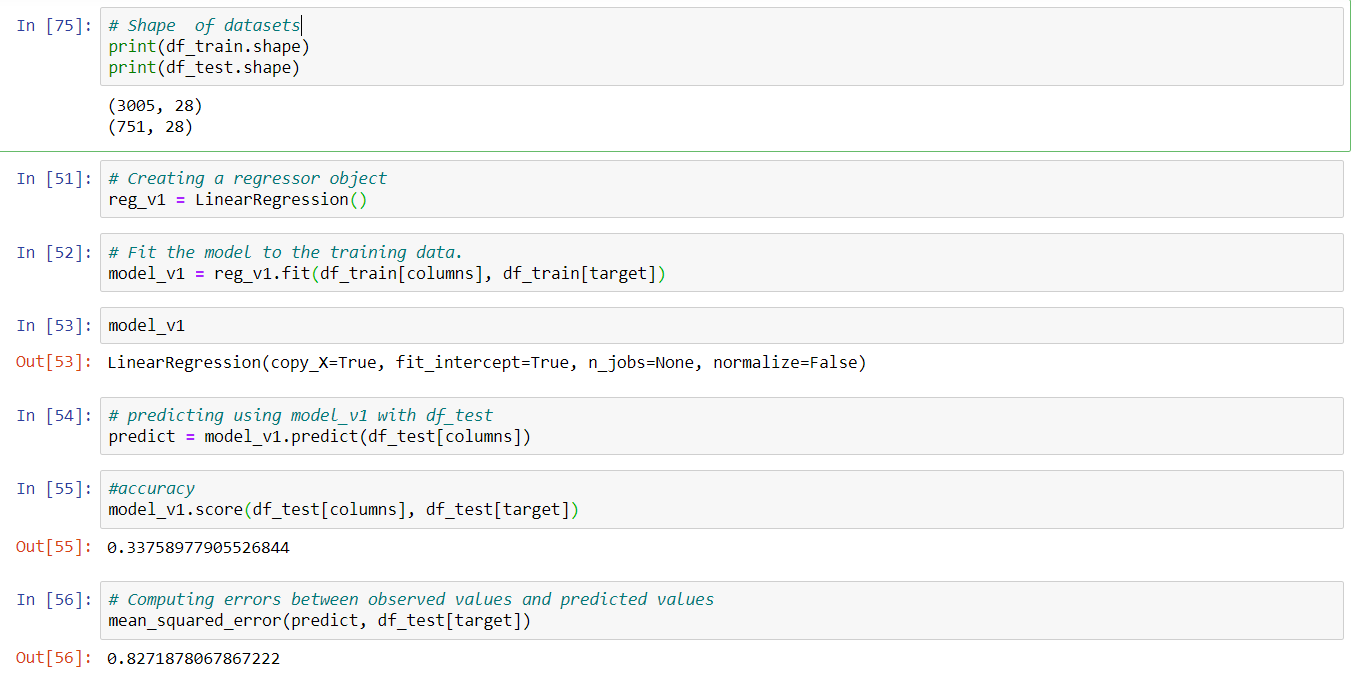
We are interested in building a model that will predict the IMDB score of

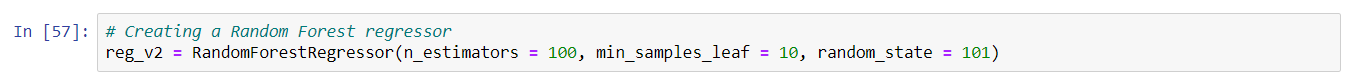
each film with a reasonable level of accuracy.

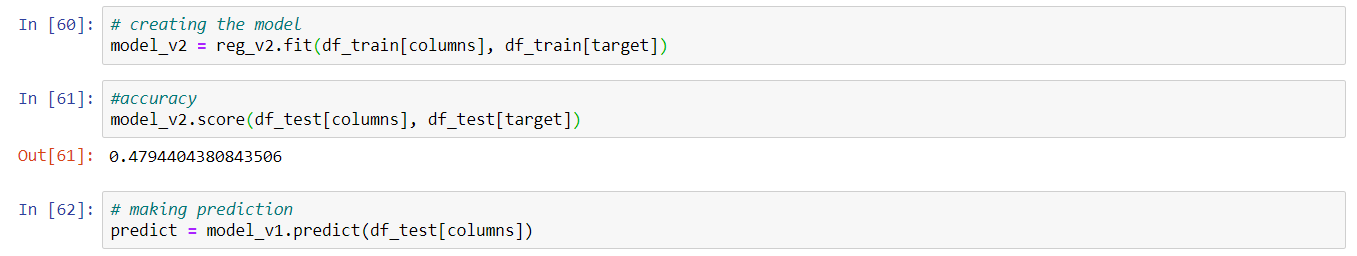
We need to split the dataset into a train subset and test subset, in other to check the accuracy of the model. 



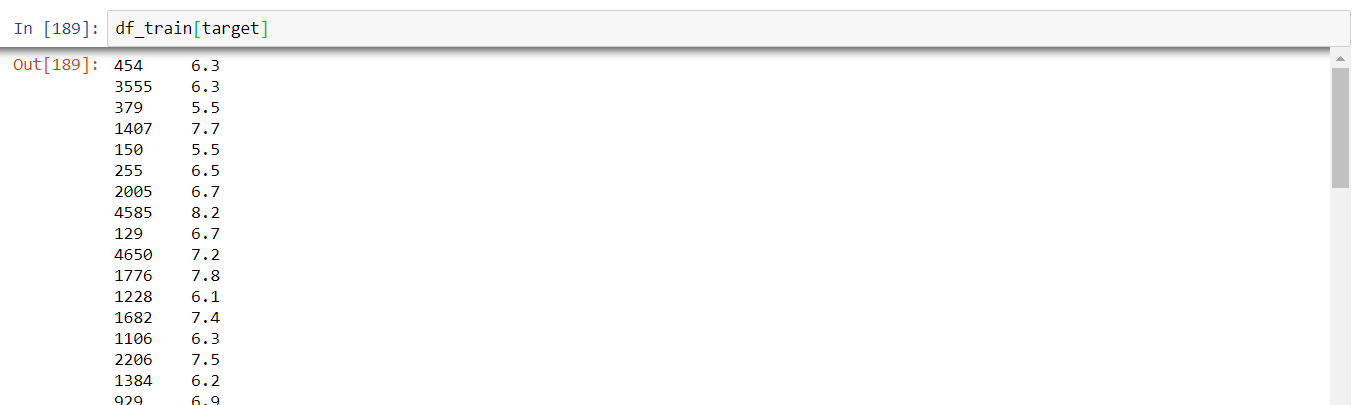
I am using two algorithms to train two models, the first is Linear regression and the second is random forest.

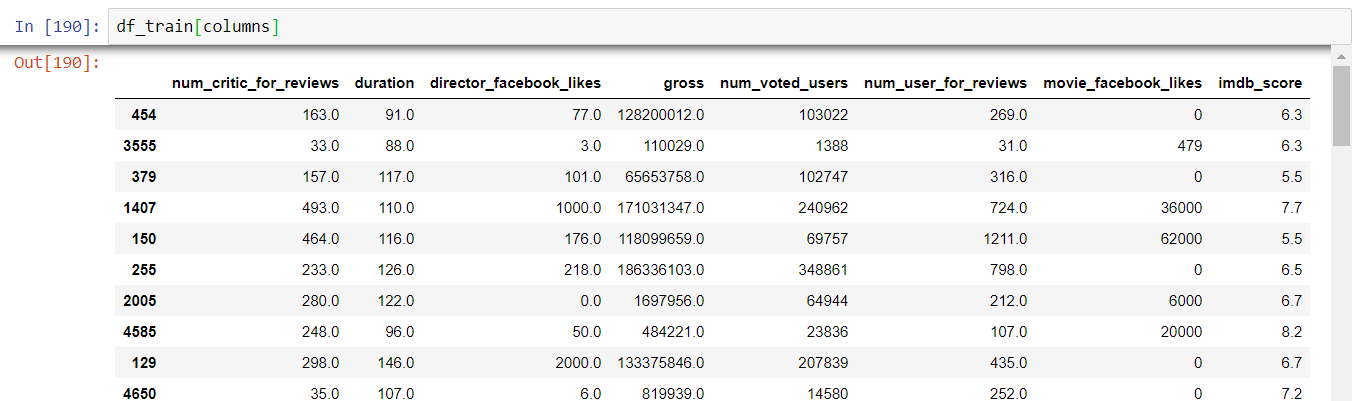


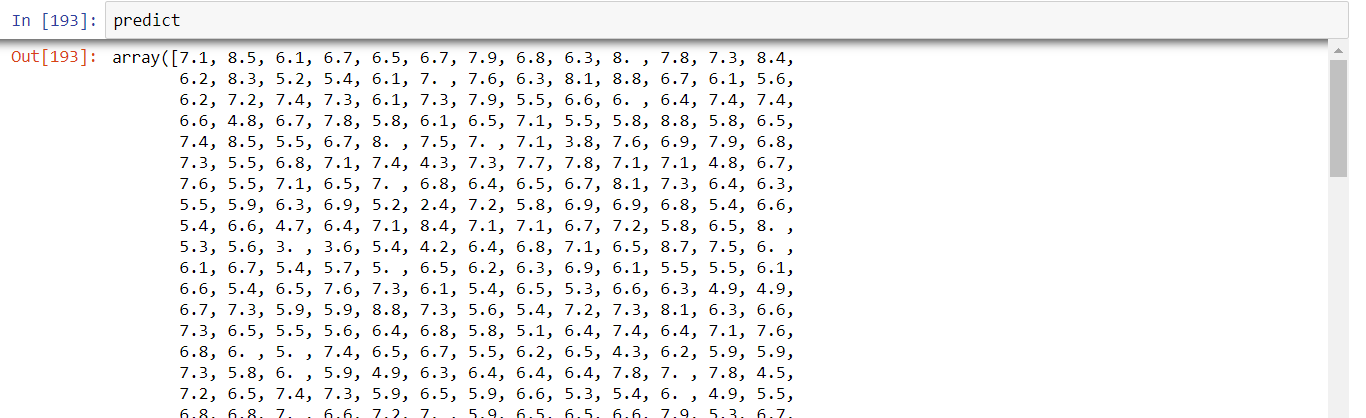












Making a test in model 1 and model 2

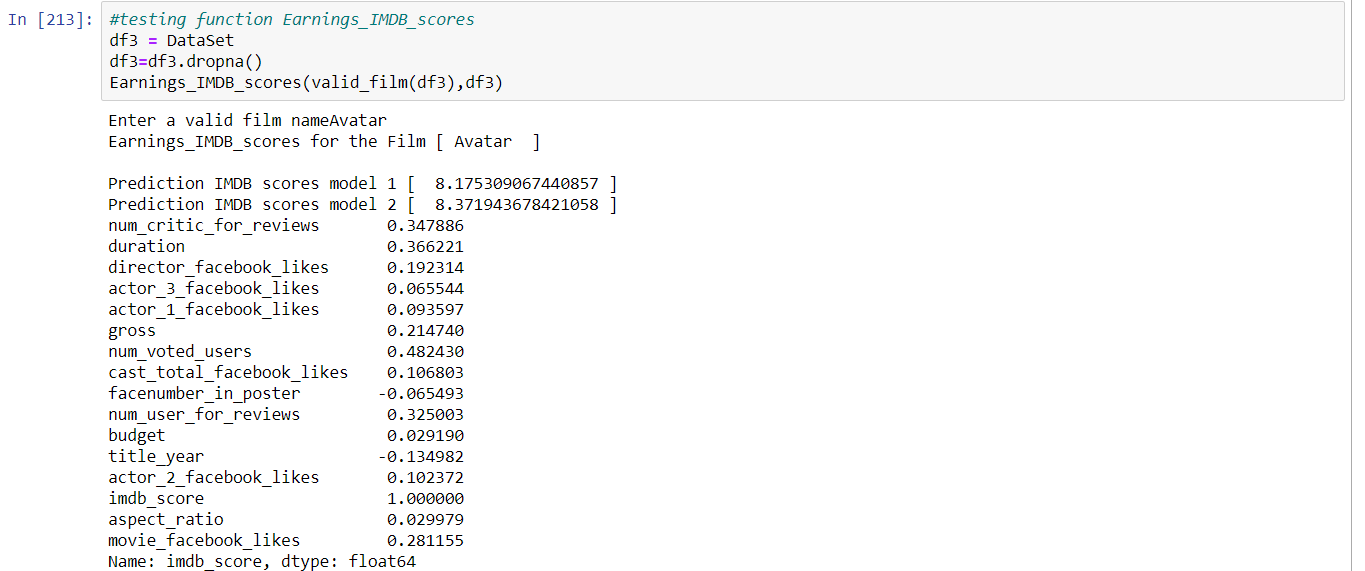
Here we can see an imdb score prediction for the selected movie.

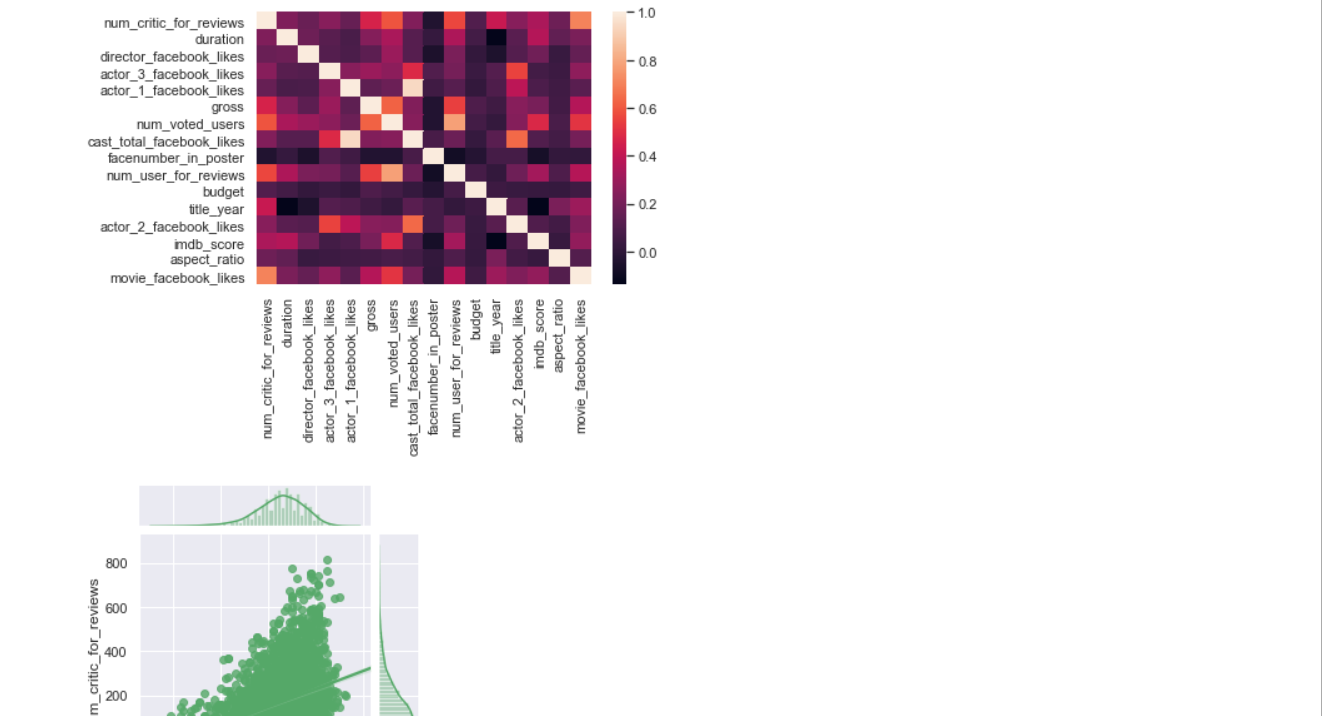


This is the function where the user will select a movie to predict imdb score.



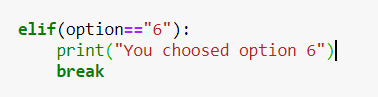
Below we can see a test of this function

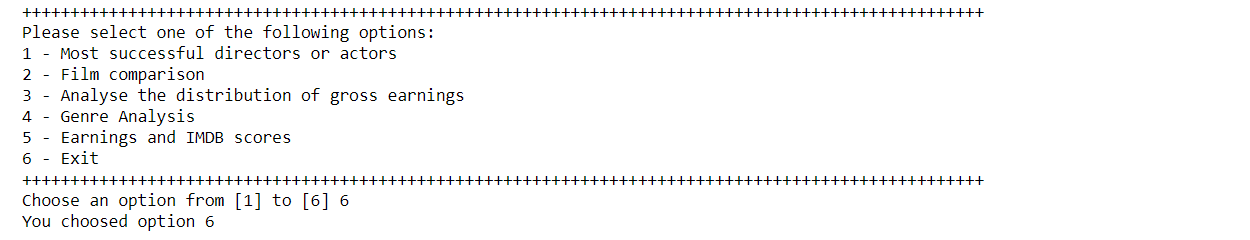




**6. Menu Option 6 – Exit**

Option 6 will “break” the loop and exit the user from the main menu.





This is the main menu, where the user can check all functions of the program.

It is a while loop that will call all functions repeatedly while the user does not choose option 6 to exit. 