**IMDB Movie Analysis**

**--By Sagir Mehmood**

**Kindly download my Jyputer Notebook:**

[**https://colab.research.google.com/drive/159hm2tehbj8bjAfWc4Ams73ETLaZC6TE?usp=sharing**](https://colab.research.google.com/drive/159hm2tehbj8bjAfWc4Ams73ETLaZC6TE?usp=sharing)

**Project Description:**

This dataset having various columns of different IMDB Movies. I have to use my knowledge of statistics and use different formulas in excel and draw necessary conclusions about the data.

**Approach:**

I analyzed it in the manner described below:

* Data cleaning
* EDA
* Analysis and conclusion

**Tech Used:** Here I sought assistance from Python Jupyter Notebook.

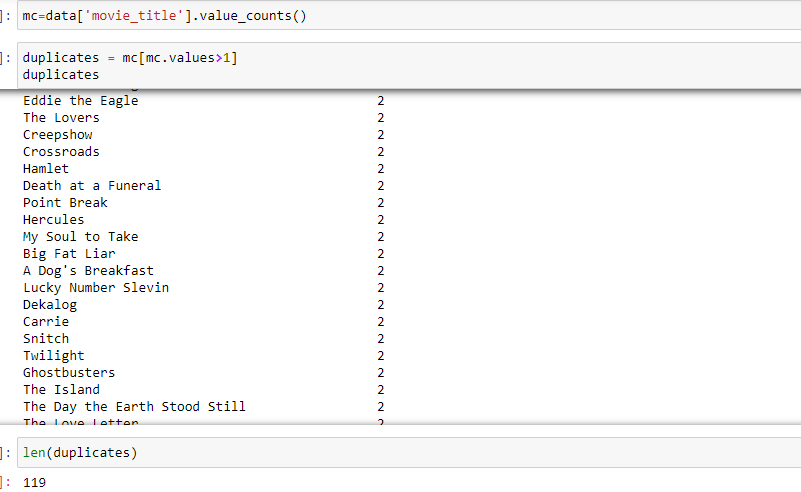
**Insights & Results:**

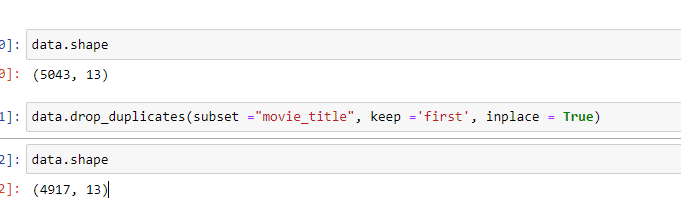
**QA. Data Cleaning:**

1. Dropping unnecessary columns (Not needed for this project):

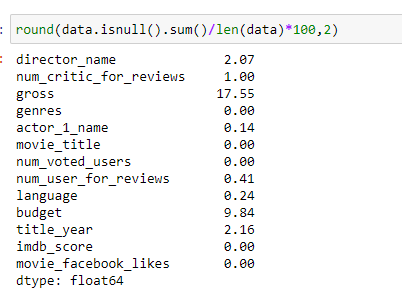
'color','director\_facebook\_likes', 'actor\_1\_facebook\_likes', 'actor\_2\_facebook\_likes', 'actor\_3\_facebook\_likes', 'actor\_2\_name', 'cast\_total\_facebook\_likes', 'actor\_3\_name', 'duration', 'facenumber\_in\_poster', 'content\_rating', 'country', 'movie\_imdb\_link', 'aspect\_ratio', 'plot\_keywords'

2. Removing the duplicate movie names, which are entered more than once.



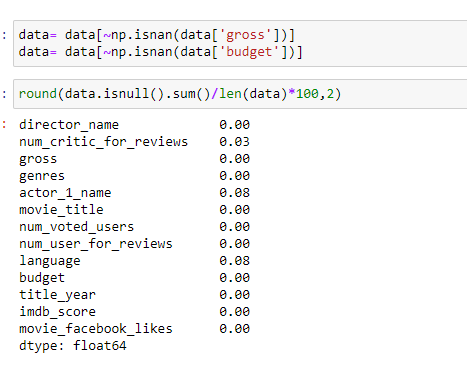


3. Checking for null values:

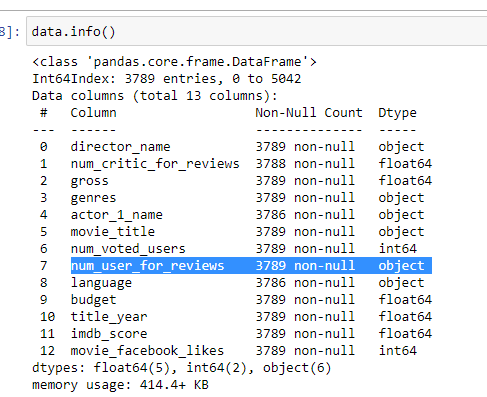


Here 'gross' and 'budget' columns have a large percentage (greater than 5%) of Null values. Here I dropped all the null rows in these two columns. Left are negligible.

After cleaning:

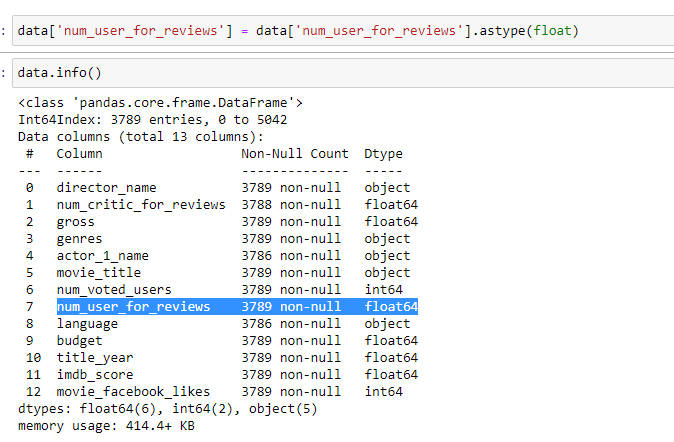


4. Data formatting:



Here num\_user\_for\_reviews column’s data are stored in object format, and I converted it to float type.

After converting:

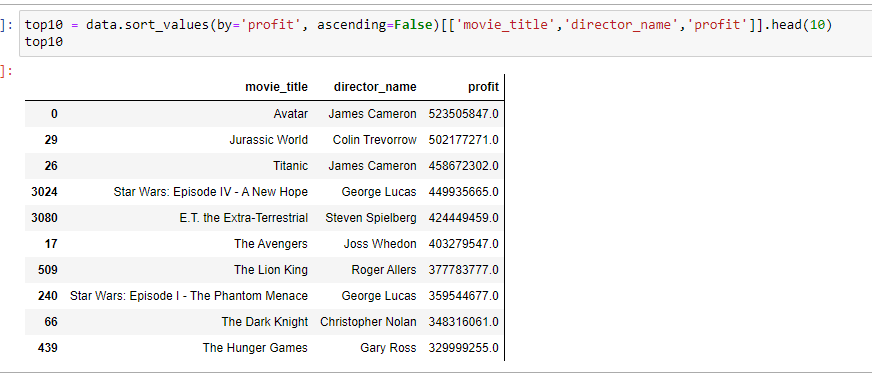


**B.**

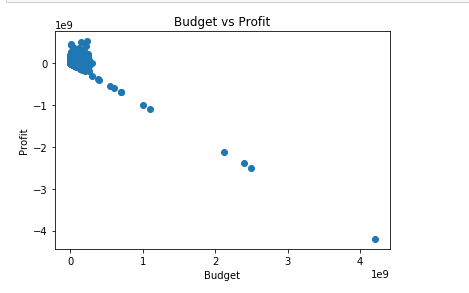
**Movies with the highest profit:**

Profit = Gross – Budget



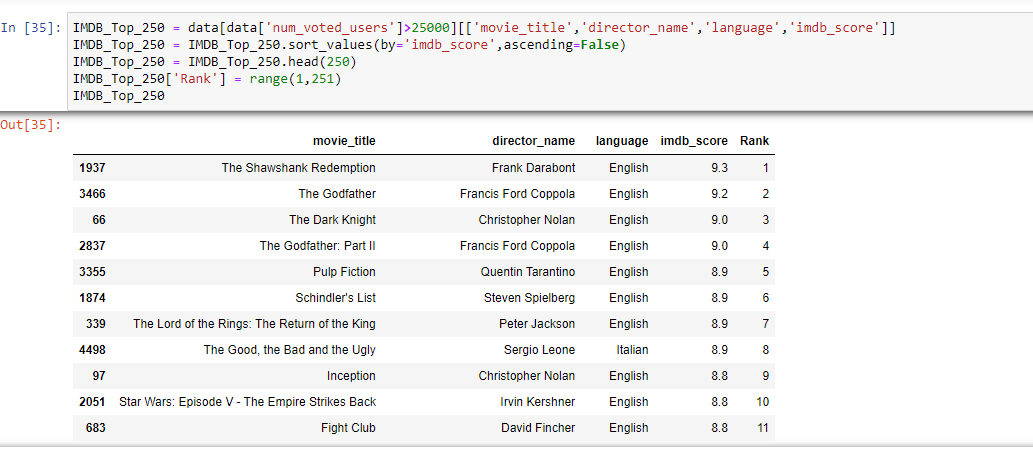
Here to 10 highest-earned movies are 

**Plot profit (y-axis) vs budget (x-axis)**

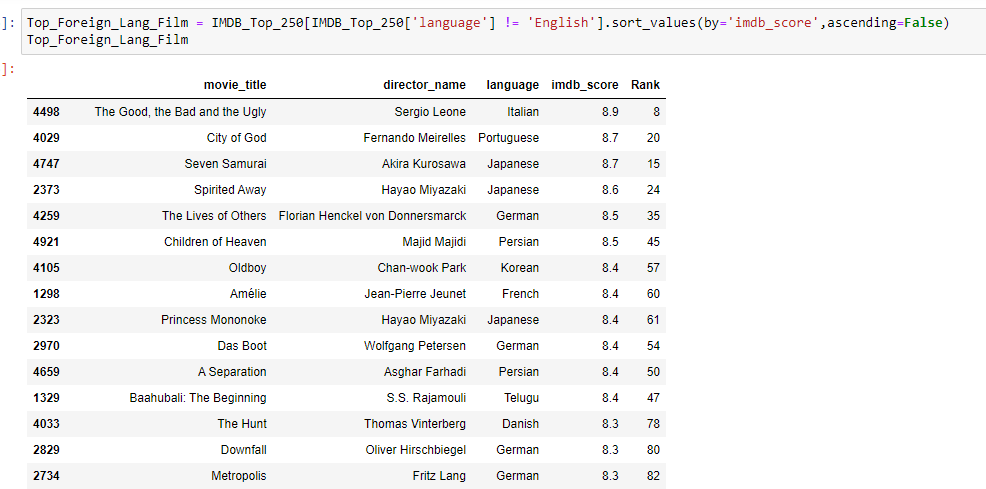


**C.**

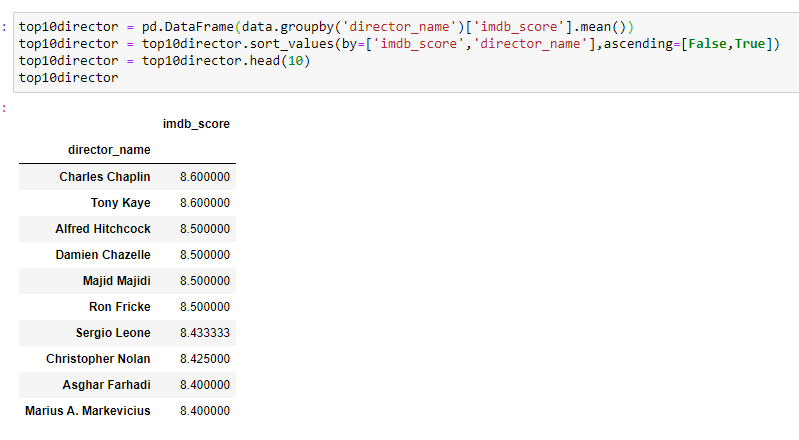
**Top 250 movies with the highest IMDb Rating for all of these movies, the num\_voted\_users is greater than 25,000:**



**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:**

****

**D. Best Director (TOP-10):**

****

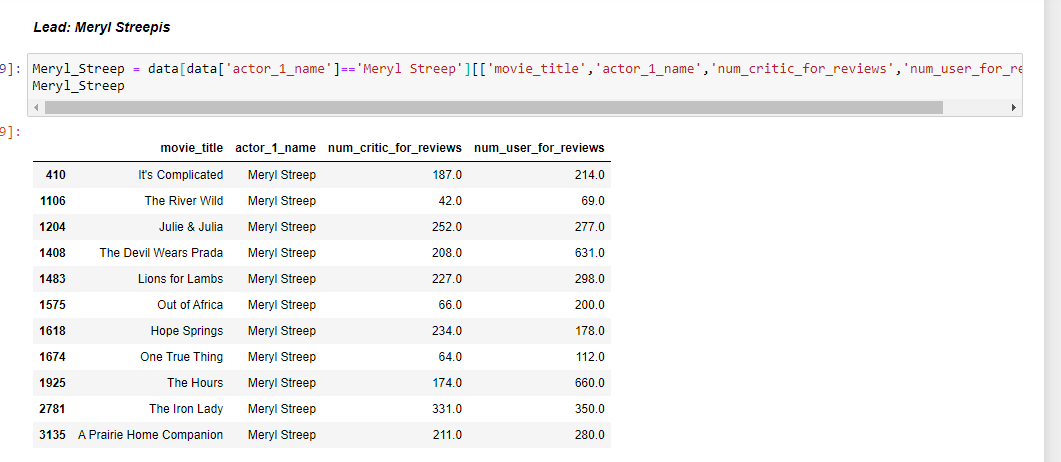
**E. Popular Genres:**

****

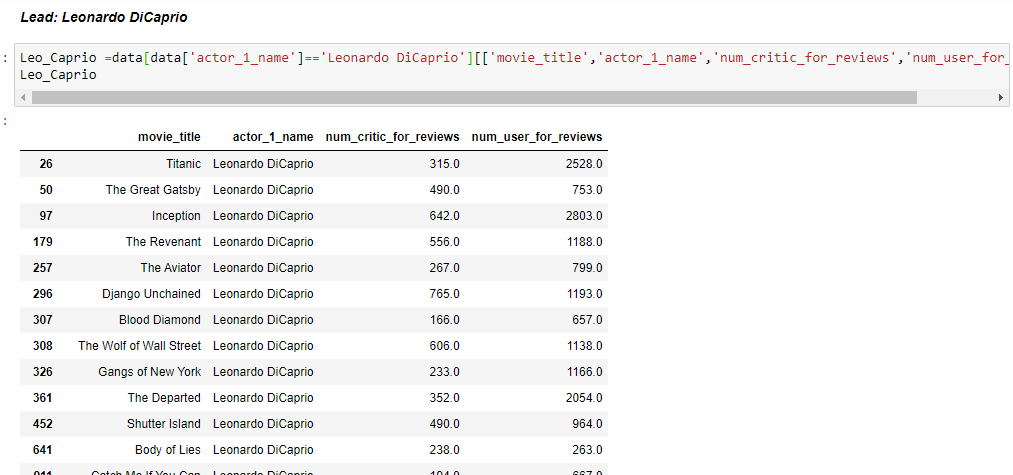
**F. Charts:**

**Movies by Leonardo DiCaprio, Brad Pitt, and Meryl Streep:**

##### Lead: Meryl Streep

****

##### Lead: Leonardo Di Caprio

****

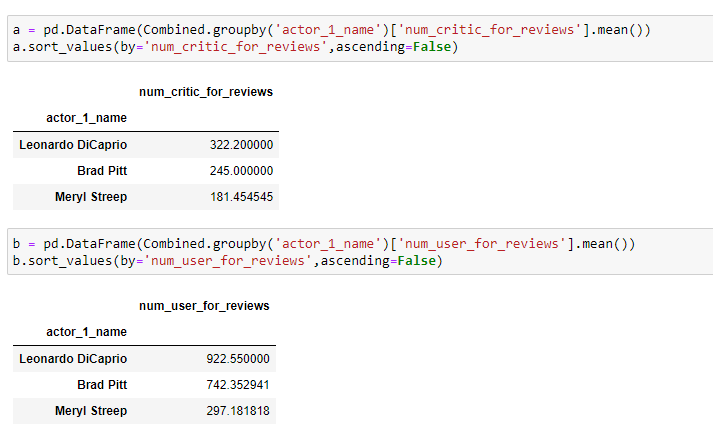
##### Lead: Brad Pitt

##### 

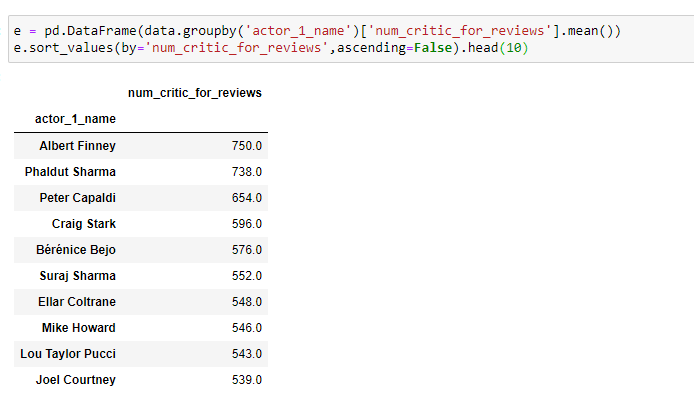
##### COMBINED:

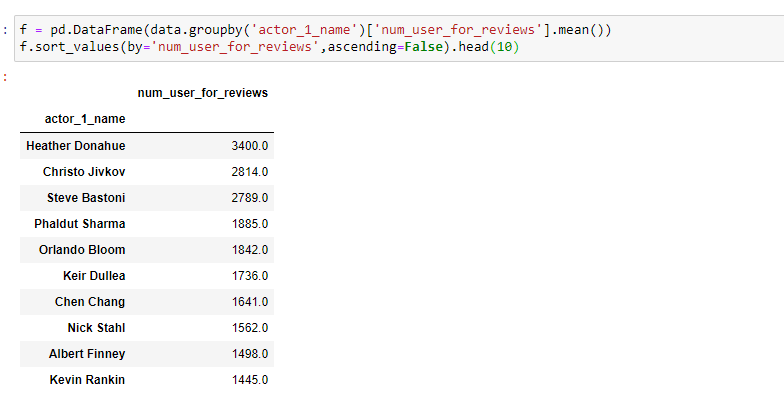
##### 

**The mean of critic reviews and audience reviews from the above data:**

****

**The mean of critic reviews and audience reviews from overall data:**

****

****

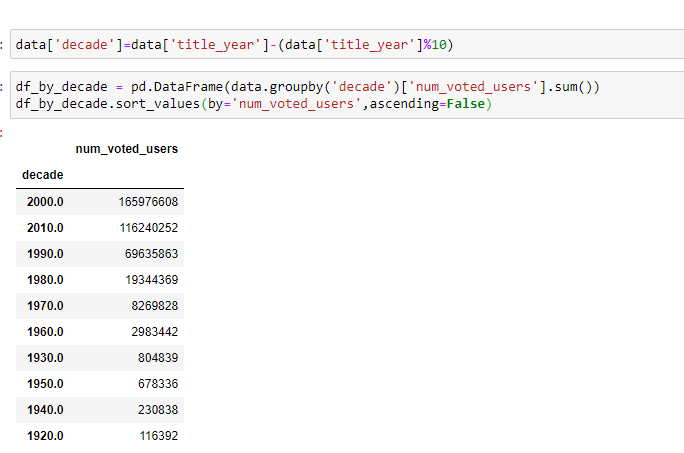
**The sum of users voted in each decade:**

Decade= year- remainder(decade/10)

Eg. Year= 2013

So, Remainder(2013/10) = 3

So, Decade = 2013-3 = 2010

****