IMDB Movie Analysis

Project Description:

Movie Industry is a very vast industry in which a lot of revenue can be generated. There are many platforms that provide movie services like Netflix, Amazon. These services also take user and critics reviews for the benefit of other users and provide movie recommendations based on the likings and disliking of the user. The data set provided for Analysis contains 45044 rows and 10 columns with columns such as director name, number of user reviews, number of critics reviews, actor 1 name, actor 2 name, title year, number of voted users, imdb score etc.

As a data analyst, I have to observe the data and try to raise a hypothesis. Also, I am supposed to draw insights from the data.

Approach:

Step 1: The data is observed thoroughly and a problem statement is to be derived. The following questions are to be asked to derive at the problem statement.

- What do you see happening?
- What is your hypothesis for the cause of the problem? (this will be broadly based on intuition initially)
- What is the impact of the problem on stakeholders?
- What is the impact of the problem not being solved?

Also, a root cause Analysis is to be performed on the data.

The following questions are to be answered:

- Find the movies with the highest profit?
- Find IMDB Top 250
- Find the best directors
- Find popular genres
- Find the critic-favourite and audience-favourite actors

Step 2: The questions are thoroughly read and only that data which is useful is put in the dataset, thus removing the unimportant columns.

Step 3: Handling Outlier Data.

Exploratory data Analysis is performed to see the outliers and missing data. There are two methods to detect outliers, Z-score method and Interquartile range method. When the distribution is perfectly uniform, Z-score method is used. However, in the numerical columns which are gross and budget, distplot is used to check the distribution of the plot. However, the distribution is not normal and there for IQR was selected to be used. However, upon using boxplot on the same columns one could observe that majority of the data is in the outlier section and could not be removed since it has high imdb score, and this data is very important.

Step 4: Handling Missing Data

Various methods have been observed using density plots and it has been concluded that mean/median imputation is best suitable and therefore applied on the columns gross and budget.

Step 5: Analysing the dataset for the problem statement

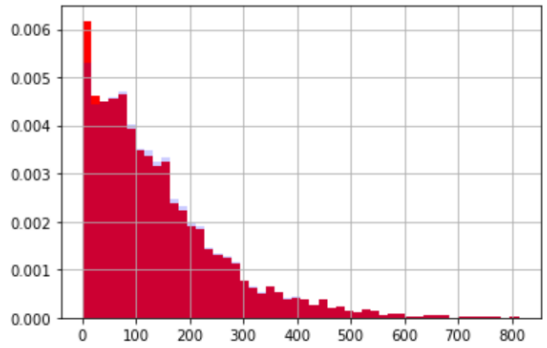
A correlation is checked among the columns and is found that the number of critics, number of voted users and number of user reviews have contributed positively to the IMDB score. However, there is a strong negative correlation with the budget. The lesser the budget, the more is the profit.

Step 6: Why Mean/Median Imputation

The median is applied to null values and is observed that the density graph is not much changing and therefore the mean/median imputation is used to handle missing data in number of critic views and number of user views.

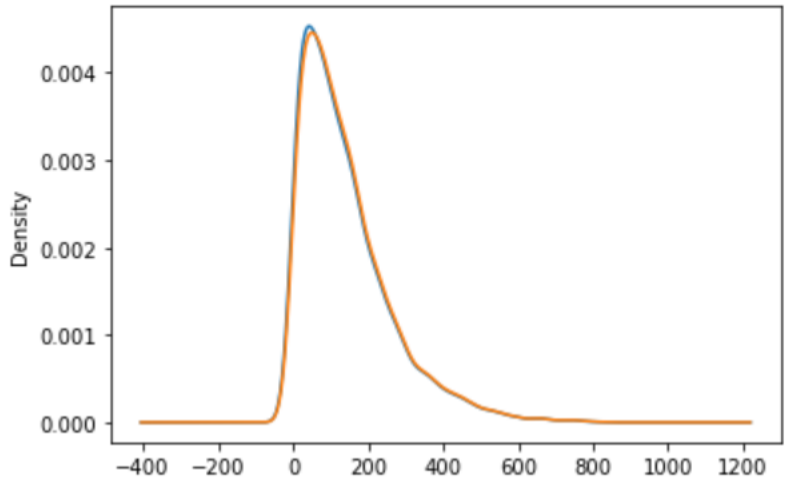
```
fig = plt.figure()
ax = fig.add_subplot(111)
movies['num_critic_for_reviews'].hist(bins = 50, ax = ax, density = True, color = 'red')
new_movies['num_critic_for_reviews'].hist(bins = 50, ax = ax, density = True, color = 'blue',alpha = 0.2)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x22f3dc498b0>
```



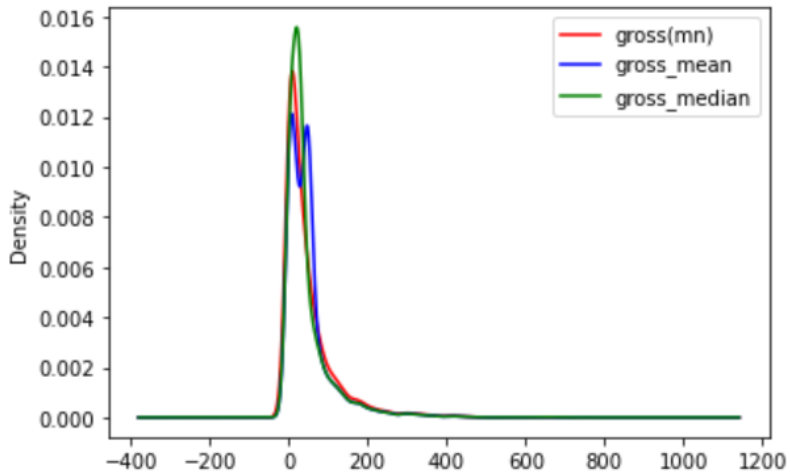

```
movies['num_critic_for_reviews'].plot.density()
new_movies['num_critic_for_reviews'].plot.density()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x22f3d794940>
```



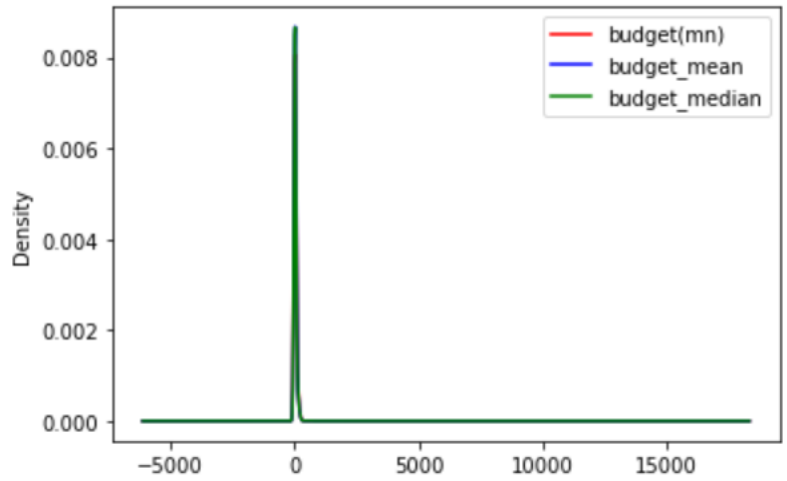
```
cols_for_imputation['gross(mn)'].plot(kind='kde',color = 'red')
cols_for_imputation['gross_mean'].plot(kind='kde',color = 'blue')
cols_for_imputation['gross_median'].plot(kind='kde',color = 'green')
plt.legend()
```

```
<matplotlib.legend.Legend at 0x22f3d8db580>
```



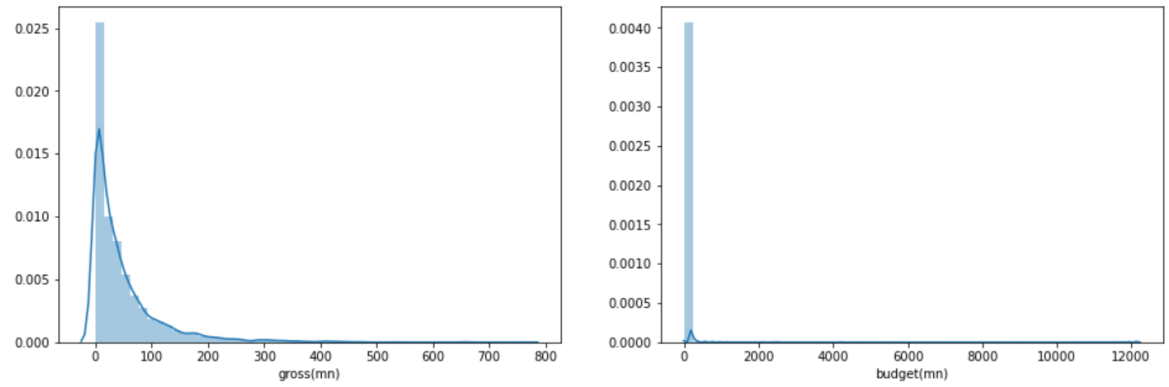

```
cols_for_imputation['budget(mn)'].plot(kind='kde',color = 'red')
cols_for_imputation['budget_mean'].plot(kind='kde',color = 'blue')
cols_for_imputation['budget_median'].plot(kind='kde',color = 'green')
plt.legend()
```

```
<matplotlib.legend.Legend at 0x22f3df7bb80>
```

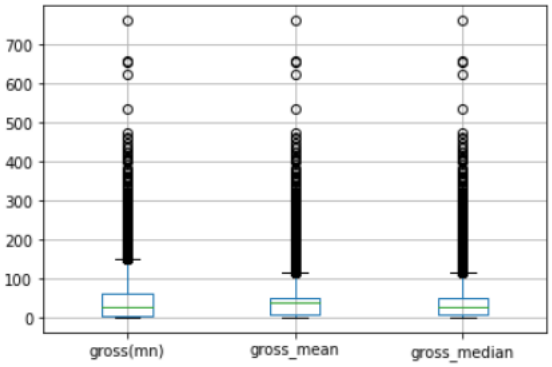


The following graphs show that the uniform distribution is not there and therefore z-score and IQR method is used to handle outliers. Also, boxplot can be used to check the amount of outliers present.

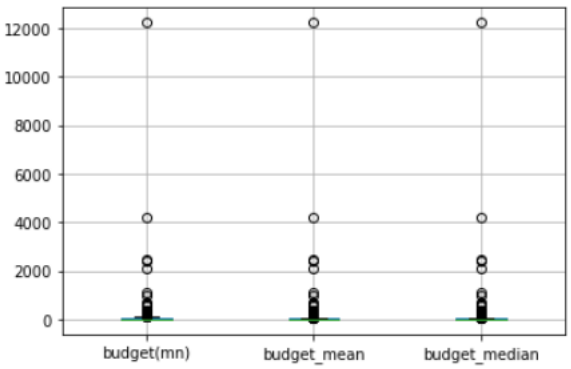
```
plt.figure(figsize = (16,5))
plt.subplot(1,2,1)
sns.distplot(cols_for_imputation['gross(mn)'])

plt.subplot(1,2,2)
sns.distplot(cols_for_imputation['budget(mn)'])
plt.show()
```




```
In [74]: cols_for_imputation[['gross(mn)','gross_mean','gross_median']].boxplot()
Out[74]: <matplotlib.axes._subplots.AxesSubplot at 0x22f3d2a9670>
```




```
In [75]: cols_for_imputation[['budget(mn)','budget_mean','budget_median']].boxplot()
Out[75]: <matplotlib.axes._subplots.AxesSubplot at 0x22f3c116d30>
```



The 12000 valued outlier in budget is removed but the other outliers seem important in terms of imdb score and are therefore capped using the Median value.

Tech Stack Used:

1. Google Sheets: Sheets are used to make graphs, pivot tables etc to draw some insights.

2. Google, YouTube and StackOverflow to clear some doubts.
3. Jupyter Notebook to perform Exploratory Data Analysis.

Insights:

1. This is one of the most important step to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)
   Your task: Clean the data

   The data Cleaning is performed and facebook likes columns are removed. After this, the following dataset is produced

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

```
movies = pd.read_csv(r"C:\Users\Swetha G\Downloads\IMDB_Movies - Practice Sheet 2.csv")
```

```
movies.head(5)
```

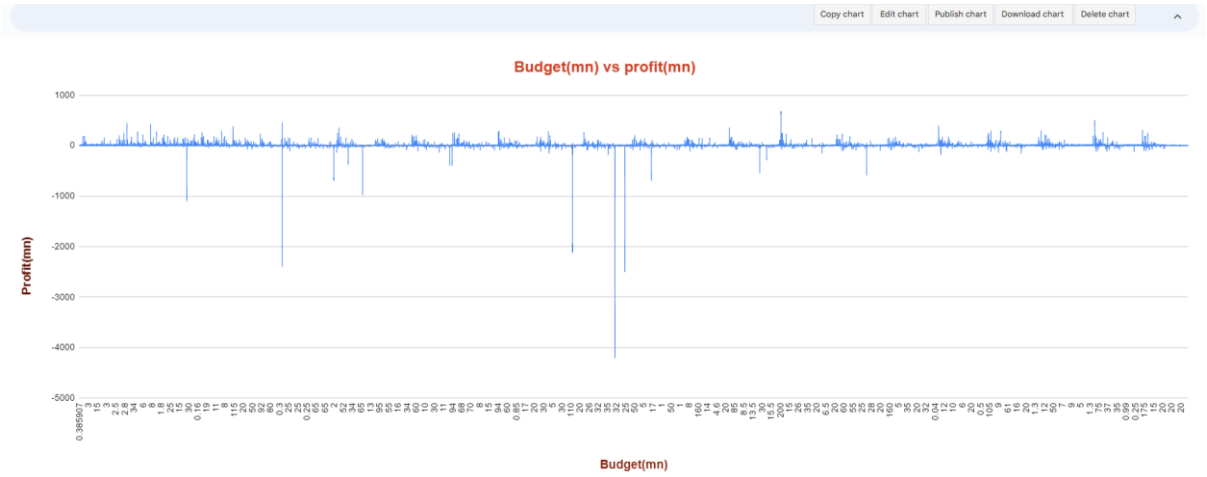
	color	director_name	duration	actor_1_name	actor_2_name	actor_3_name	budget(mn)	gross(mn)	profit(mn)	genres	movie_title	num
0	Color	James Cameron	178.0	CCH Pounder	Joel David Moore	Wes Studi	75.28	760.50	685.22	Action|Adventure|Fantasy|Sci-Fi	Avatar	
1	Color	James Cameron	194.0	Leonardo DiCaprio	Kate Winslet	Gloria Stuart	200.00	658.67	458.67	Drama|Romance	Titanic	
2	Color	Colin Trevorrow	124.0	Bryce Dallas Howard	Judy Greer	Omar Sy	150.00	652.17	502.17	Action|Adventure|Sci-Fi|Thriller	Jurassic World	
3	Color	Joss Whedon	173.0	Chris Hemsworth	Robert Downey Jr.	Scarlett Johansson	220.00	623.27	403.27	Action|Adventure|Sci-Fi	The Avengers	
4	Color	Christopher Nolan	152.0	Christian Bale	Heath Ledger	Morgan Freeman	185.00	533.31	348.31	Action|Crime|Drama|Thriller	The Dark Knight	

```
movies.isnull().mean()*100
```

```
color                      0.360360
director_name              2.062062
duration                   0.280280
actor_1_name               0.140140
actor_2_name               0.260260
actor_3_name               0.440440
budget(mn)                 9.709710
gross(mn)                 18.398398
profit(mn)                 0.000000
genres                     0.000000
movie_title                0.000000
num_voted_users            0.000000
num_critic_for_reviews     0.960961
num_user_for_reviews       0.400400
language                   0.220220
title_year                 2.122122
imdb_score                 0.000000
dtype: float64
```

   The number of null values is calculated and accordingly missing data and outliers can be modified. Mean/Median Imputation is used to fill the missing data.

2. Movies with highest profit: Create a new column called profit which contains the difference of the two columns: gross and budget. Sort the column using the profit column as reference. Plot profit (y-axis) vs budget (x- axis) and observe the outliers using the appropriate chart type.
Your task: Find the movies with the highest profit?



A new column is created and the profit column is plotted against the budget column.

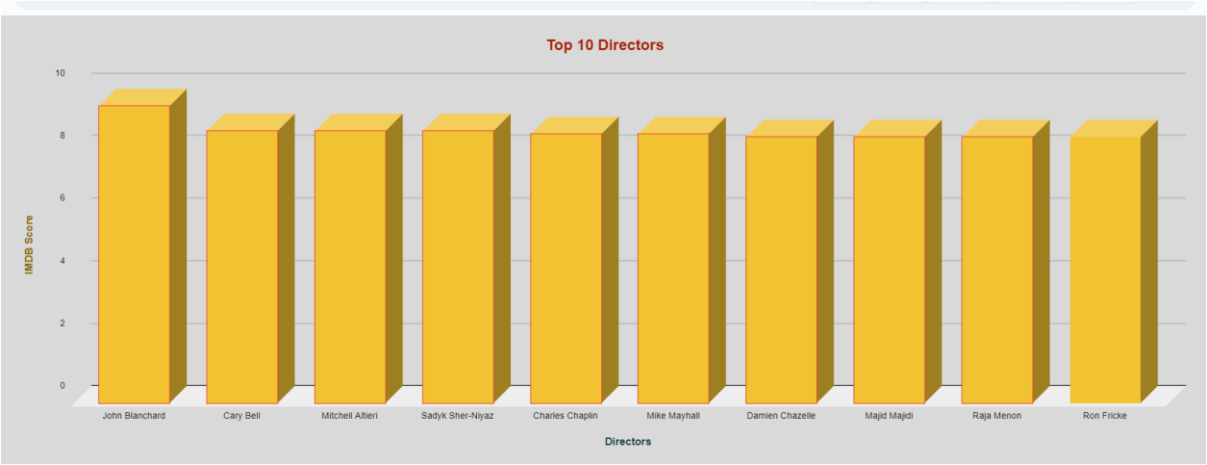
3. Top 250: Create a new column IMDb_Top_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb_score). Also make sure that for all of these movies, the num_voted_users is greater than 25,000. Also add a Rank column containing the values 1 to 250 indicating the ranks of the corresponding films.

   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. You can use your own imagination also!
   Your task: Find IMDB Top 250

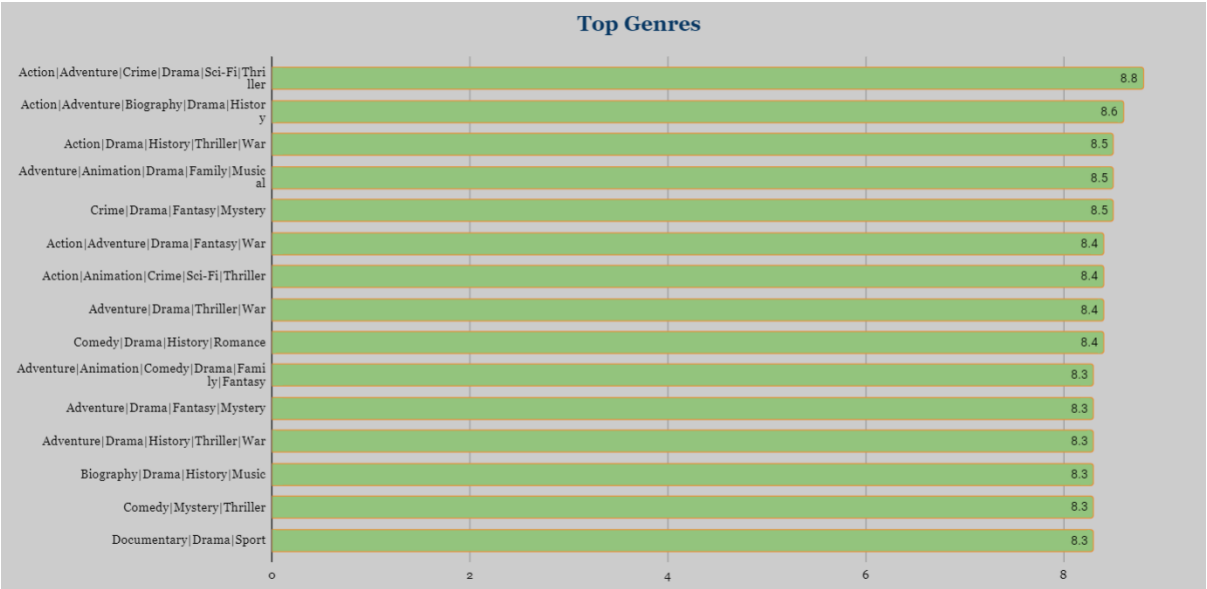
	Top_250_and_voted	Top_Foreign_Lang_Film
1	12 Angry Men	×
2	12 Years a Slave	×
3	2001: A Space Odyssey	×
4	A Beautiful Mind	×
5	A Christmas Story	×
6	A Separation	
7	Airlift	
8	Akira	
9	Aladdin	×
10	Aliens	×
11	Alien	×
12	Amadeus	×
13	American Beauty	×
14	American History X	×
15	Amores Perros	
16	Amélie	
17	Annie Hall	×
18	Apocalypse Now	×
19	Baahubali: The Beginning	
20	Back to the Future	×
21	Barry Lyndon	×
22	Batman Begins	×
23	Batman: The Dark Knight Returns, Part 2	×
24	Before Sunrise	×

4. Best Directors: Group the column using the director_name column.

   Find out the top 10 directors for whom the mean of imdb_score is the highest and store them in a new column top10director. In case of a tie in IMDb score between two directors, sort them alphabetically.
   Your task: Find the best directors



5. Popular Genres: Perform this step using the knowledge gained while performing previous steps.
   Your task: Find popular genres



6. **Charts:** Create three new columns namely, Meryl_Streep, Leo_Caprio, and Brad_Pitt which contain the movies in which the actors: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors. Use only the actor_1_name column for extraction. Also, make sure that you use the names 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' for the said extraction.
   You Task: Find the mean of the num_critic_for_reviews and num_users_for_review and identify the actors which have the highest mean.

Maximum User reviews for

Heather Donahue

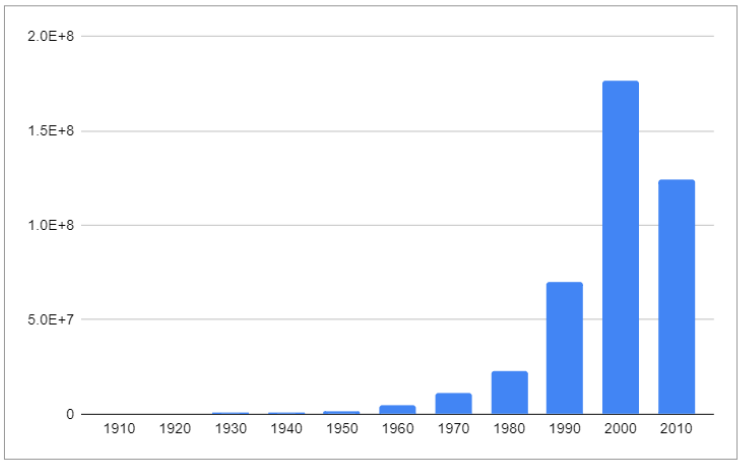
Maximum Critic reviews for

Phaldut Sharma

7. Observe the change in number of voted users over decades using a bar chart. Create a column called decade which represents the decade to which every movie belongs to. For example, the title_year year 1923, 1925 should be stored as 1920s. Sort the column based on the column decade, group it by decade and find the sum of users voted in each decade. Store this in a new data frame called df_by_decade.

Your task: Find the num_of_voted users per decade

decade	SUM of num_voted_users
1910	10718
1920	128672
1930	984397
1940	1211888
1950	1638504
1960	5065074
1970	11072372
1980	23021112
1990	70278714
2000	176161342
2010	124382231



Result:

1. It is observed that the summarising and cleaning the data is the most important step in drawing good insights from the data.
2. The budget corresponding to the most profitable movie is 25 million.
3. A list of Top 250 movies is given with “The Shawshank Redemption” topping the list. “The Good, the bad and the ugly” is the top foreign Language film.
4. John Blanchard is the top director
5. Action|Adventure|Crime|Drama|Sci-Fi|Thriller is the popular genre.
6. Heather Donahue and Phaldut Sharma top the list in User reviews and Critic review respectively/
7. Maximum number of people voted in the year 2000.