A War Through The 1000 Best Movies

Editar artículo
Ver estadísticas

Juan Archidona Ahijado
Data Scientist

1 artículo

1 de septiembre de 2022

Over the years, cinema has been one of the arts most conditioned by social, economic and technological changes. Until a few years ago, some production studios and film makers sought to achieve a certain artistic sense, along with a good script and solid performances, in order to provide us with a great experience that would last over time.

Many moviegoers have seen a clear shift in the planning, production and marketing of films in recent years. Nowadays, in most cases, films are conceived as products for immediate consumption, tailored to the target audience and seeking to achieve the highest possible profit.

And we can say that superhero movies are the maximum exponent of this trend. Different production studios, directors and actors have embarked on this profitable way of making movies, but not all of them are happy with it...

And Scorsese took the floor...

Martin Scorsese, one of the most prestigious and acclaimed directors in history, had something to say about it: in

October 2019 he told Empire magazine (* **source**) that he had tried to watch some Marvel Studios superhero movies, had not enjoyed them very much and had decided that "they were not cinema".

With that he could be satisfied, but he moved on. For Scorsese, these films were based on a crude impersonality. "Honestly, the closest thing I can think of, however well made they are, with actors doing the best they can given the circumstances, are theme parks," he told Empire. "It's not the cinema of human beings trying to convey emotional and psychological experiences to another human being."

It should be recalled that Marvel Studios released its first movie "Iron Man" in 2008. Since then, they have reaped a lot of box office successes and influenced in different ways the critics and tastes of viewers.

But Scorsese's criticisms are especially relevant again now. It's a little strange to say, but Marvel is in its failing season. Or, at least, as close as Marvel can get to an era of failures:

In the past 18 months, we've had six Marvel Studios movies and seven TV series, with notably declining box office and positive reviews; in the next 18 months, there are another four movies, four series and two specials planned.

With such a crowded calendar, each new Marvel Studios announcement seems harmless. Moreover, the gradual loss of characters that even people who went to the movies twice a year knew - Iron Man, Captain America, Black Widow - makes the lack of direction even more glaring.

Scorsese has maintained his stance on superhero movies in different interviews. Being supported by a good part of the public and different "old-school" colleagues, like Francis Ford Coppola, who went so far as to qualify superhero movies as "despicable".

We can establish that Martin Scorsese's criticism points directly to the evolution of the cinema industry, the success of his films and their acceptance by critics and audiences.

And here several questions arise:

- Is Martin Scorsese's point of view justified?.

- Before and since Marvel Studios jumped into the field in 2008... How have the film industry, its best films, critics and

the public's opinion evolved?

- Can we compare Scorsese's and Marvel's movies?

Let's try to find the answers...

Describing our data and analysis

You can find the Python Notebook code at the following **link**

We will start by reading, cleaning and analyzing a first dataset that stores information about the top 1000 movies rated on imdb.com, with the following fields:

- Poster_Link: poster from the movie.
- Series_Title: title from the movie.
- Released_Year: year at which the movie was released.
- Certificate: a label stating who should be allowed to wa tch the movie.
- Runtime: duration of the movie.
- Genre: the different genres of the movie stored in the s ame column.
- IMDB_Rating: movie rating from IMDB.com.
- Overview: movie theme description.
- Meta_score: weighted average of reviews from metacrit ic.com
- Director: movie director.
- Star1, Star2, Star3 and Star4: main cast of the movie se parated in different columns.
- No_of_Votes: sum of votes rating the movie.
- Gross: cumulative earnings of the movie in US and Can ada.

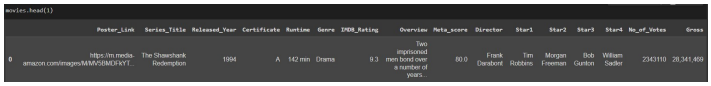
```
# import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_theme(style = "darkgrid")

# read dataset
movies = pd.read_csv("https://raw.githubusercontent.com/JuanArchidona/practica_pandas/main/imdb_top_1000.csv", sep = ",")
```

In the first instance, our main objective is to analyze the evolution of the top 1000 movies through the following parameters:

- Whether the movie was released before or since 2008, when Marvel Studios released "Iron Man".

- Cross comparisons according to ratings, gross, genre, year of release and number of votes.

Next we are going to query a record and the value types of each field




```
movies.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 16 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Poster_Link    1000 non-null   object 
 1   Series_Title   1000 non-null   object 
 2   Released_Year  1000 non-null   object 
 3   Certificate    899 non-null    object 
 4   Runtime        1000 non-null   object 
 5   Genre          1000 non-null   object 
 6   IMDB_Rating    1000 non-null   float64
 7   Overview       1000 non-null   object 
 8   Meta_score     843 non-null    float64
 9   Director       1000 non-null   object 
 10  Star1          1000 non-null   object 
 11  Star2          1000 non-null   object 
 12  Star3          1000 non-null   object 
 13  Star4          1000 non-null   object 
 14  No_of_Votes    1000 non-null   int64  
 15  Gross          831 non-null    object 
dtypes: float64(2), int64(1), object(13)
memory usage: 125.1+ KB
```

We can see that there are several fields that do not provide relevant information to our analysis. Or whose data type does not match the required one.

We proceed to adjust these fields and their values, using methods such as rename(), sort_values(), query(), replace(), assign() or drop()... This is a summary of the steps necessary to make the changes:

We need to modify the type of all "Year" values.

```
[10] movies = movies.assign(Year = lambda dataset: dataset.Year.astype(int))
```

We also need to modify the values of "Meta_Score" for later comparisons.

```
[11] movies = movies.assign(Meta_Score = lambda dataset: dataset.Meta_Score.replace(np.nan, 0).astype(int))
```

Next we are going to limit the column "Genre", in which we will store just the main genre of each movie.

```
[12] movies = movies.assign(Genre = lambda dataset: dataset.Genre.apply(lambda text: text.split(",")[0]))
```

We also need to adjust "Gross" values and type, replacing empty values.

```
[13] movies = movies.assign(Gross = lambda dataset: dataset.Gross.str.replace(",", "").astype("float64").replace(np.nan, 0).astype(int))
```

Marvel Studios began releasing its own movies in 2008 with "Iron Man".

```
[14] movies.query("Title == 'Iron Man'", engine = "python")["Year"]

     502    2008
     Name: Year, dtype: int64
```

Let's create another column called "Marvel". With which we are going to classify all the movies, depending on whether they were released before or since "Marvel" began releasing its own movies.

```
[15] movies = movies.assign(Marvel = lambda dataset: dataset.Year.map(lambda value: "Before" if value < 2008 else "Since"))
```

Finally we are going to drop some columns that do not provide relevant information to our analysis.

```
[16] movies = movies.drop(columns = {"Poster_Link", "Certificate", "Runtime", "Overview", "Star1", "Star2", "Star3", "Star4"})
```

Let's check the final format of the dataset, parameters and their values.

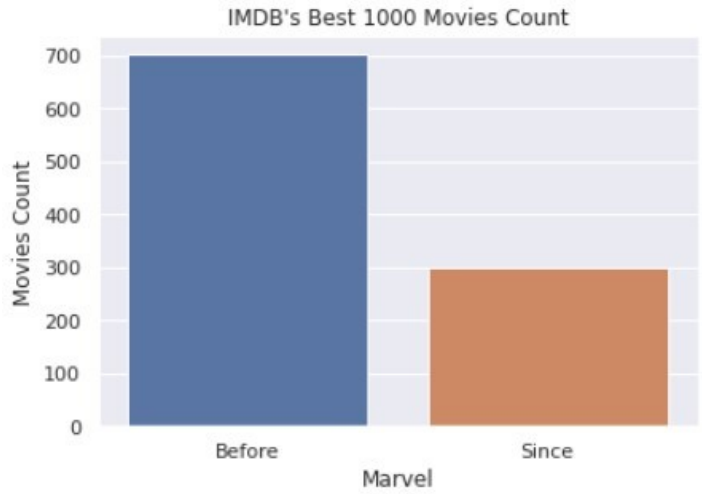
```
[17] movies.head()
```

	Title	Year	Genre	Rating	Meta_Score	Director	Votes	Gross	Marvel
0	The Shawshank Redemption	1994	Drama	9.3	80	Frank Darabont	2343110	28341469	Before
1	The Godfather	1972	Crime	9.2	100	Francis Ford Coppola	1620367	134966411	Before
2	The Dark Knight	2008	Action	9.0	84	Christopher Nolan	2303232	534858444	Since
3	The Godfather: Part II	1974	Crime	9.0	90	Francis Ford Coppola	1129952	57300000	Before
4	12 Angry Men	1957	Crime	9.0	96	Sidney Lumet	689845	4360000	Before

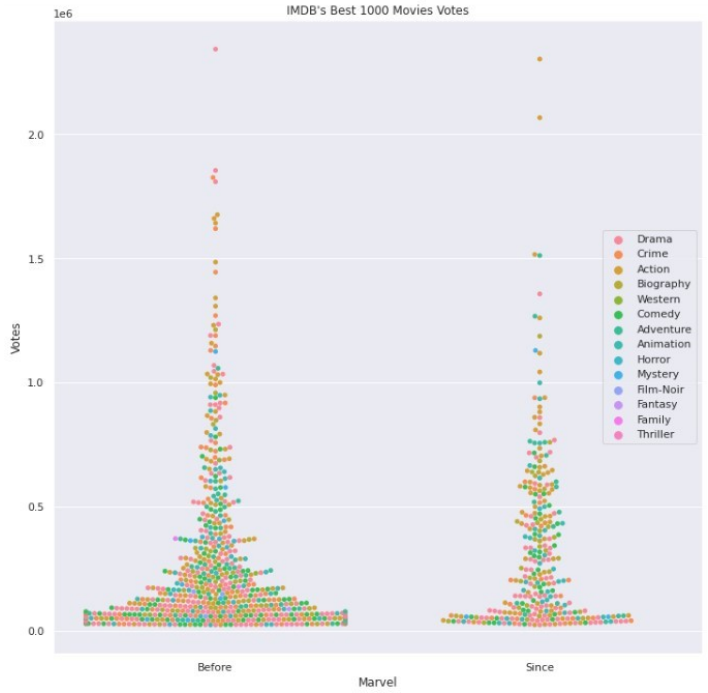
```
[18] movies.info()

     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 1000 entries, 0 to 999
     Data columns (total 9 columns):
      #   Column      Non-Null Count  Dtype
     ---  ------      --------------  -----
      0   Title       1000 non-null   object
      1   Year        1000 non-null   int64
      2   Genre       1000 non-null   object
      3   Rating      1000 non-null   float64
      4   Meta_Score  1000 non-null   int64
      5   Director    1000 non-null   object
      6   Votes       1000 non-null   int64
      7   Gross       1000 non-null   int64
      8   Marvel      1000 non-null   object
     dtypes: float64(1), int64(4), object(4)
     memory usage: 70.4+ KB
```

Values to consider

It should be noted that the number of films in the dataset prior to the landing of Marvel Studios, is higher than the number of films after. As we can see in the following graph:



In the following graph we can see the distribution of votes, according to the main genre of the film.

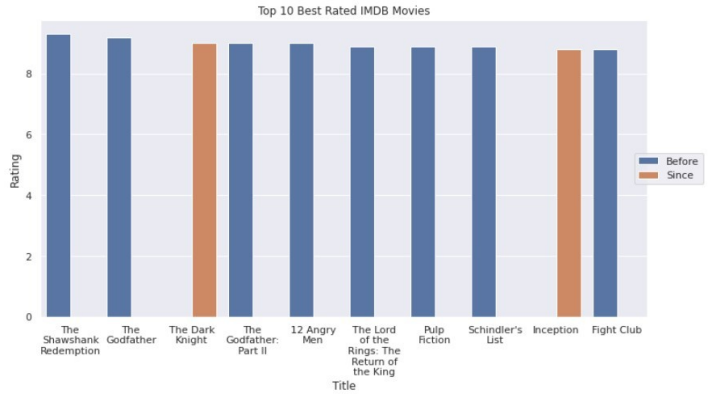


Analysis by rating

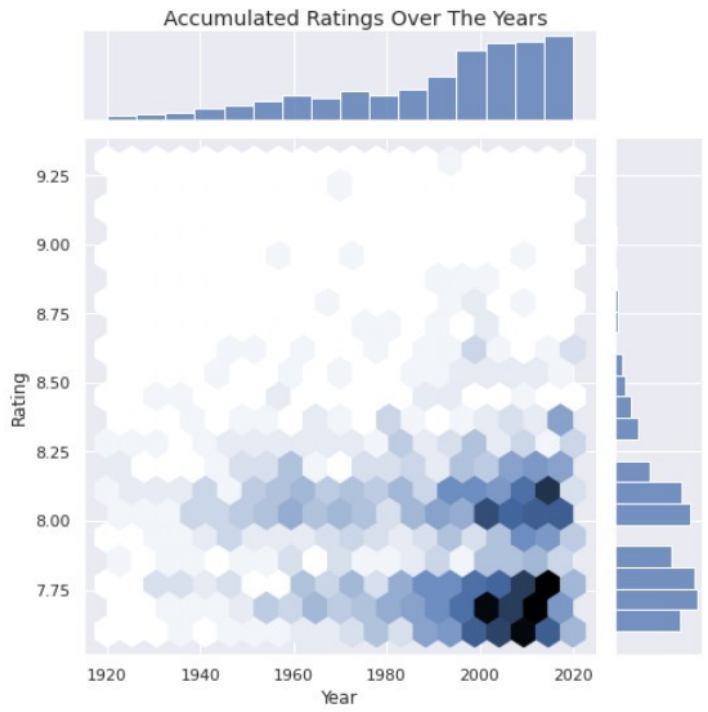
Below we will check out the top 10 highest rated movies, with some interesting detail:

```
movies.filter(["Title", "Rating", "Marvel"]).head(10)
```

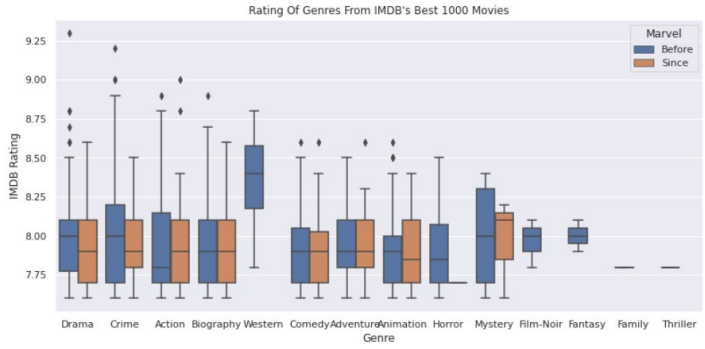
	Title	Rating	Marvel
0	The Shawshank Redemption	9.3	Before
1	The Godfather	9.2	Before
2	The Dark Knight	9.0	Since
3	The Godfather: Part II	9.0	Before
4	12 Angry Men	9.0	Before
5	The Lord of the Rings: The Return of the King	8.9	Before
6	Pulp Fiction	8.9	Before
7	Schindler's List	8.9	Before
8	Inception	8.8	Since
9	Fight Club	8.8	Before



It turns out that most of the top-rated movies according to IMDB predate the emergence of Marvel Studios. We could establish that the passage of time and popularity, do not necessarily lead to better movies.



The ratings according to the main genre of the film, show that some themes have lost relevance and acceptance, over the years, in the eyes of critics. As Martin would surely have asserted.

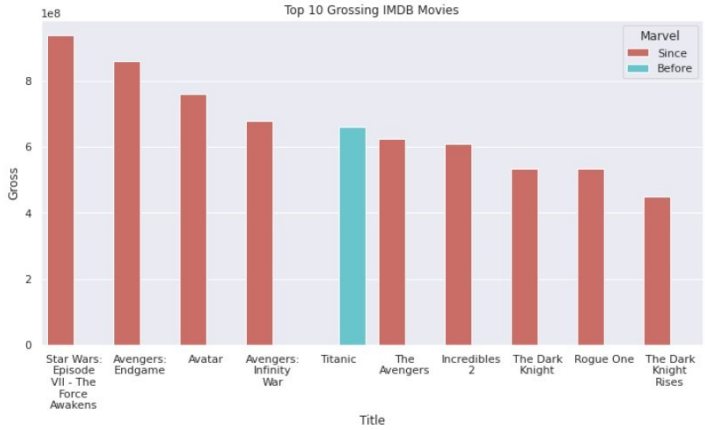


Analysis according to gross

Next, let's check out IMDB's top 10 highest grossing movies. There are also interesting details:

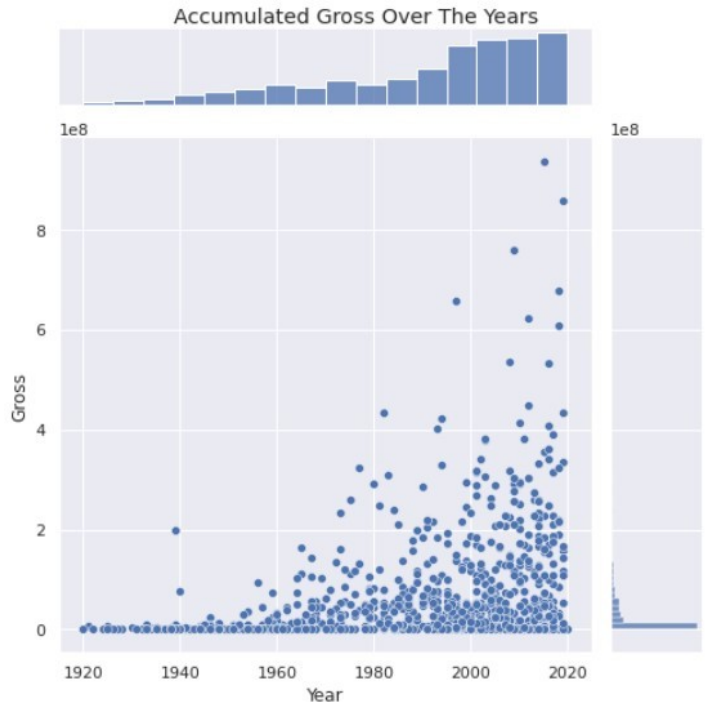
```
movies.sort_values("Gross", ascending = False).filter(["Title", "Gross", "Marvel"]).head(10)
```

	Title	Gross	Marvel
477	Star Wars: Episode VII - The Force Awakens	936662225	Since
59	Avengers: Endgame	858373000	Since
623	Avatar	760507625	Since
60	Avengers: Infinity War	678815482	Since
652	Titanic	659325379	Before
357	The Avengers	623279547	Since
891	Incredibles 2	608581744	Since
2	The Dark Knight	534858444	Since
582	Rogue One	532177324	Since
63	The Dark Knight Rises	448139099	Since



It is enlightening that 9 of the top 10 grossing films happen to be after Marvel Studios came on the scene. And that 6 of the 10 films are superhero movies. Next to 2 of Star Wars movies and Avatar, with similar subject matter.

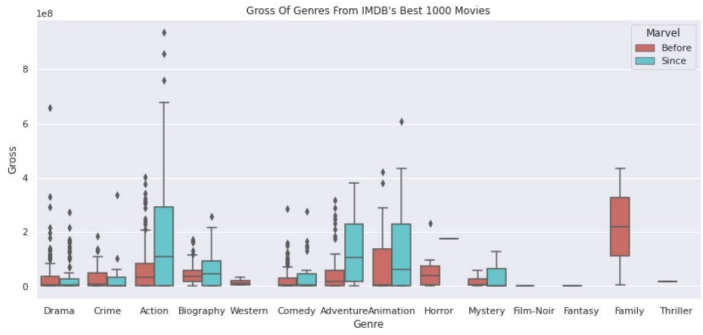
It seems that Scorsese's criticism, about the industry trend and the marked taste of the viewers, is well-founded.



The earnings of films according to their main genre show that some thematic films have achieved very high revenues

over time. Leaving other genres behind.

Martin Scorsese may still be right.



Describing more data and its analysis

We will continue by reading, cleaning and analyzing the second dataset that stores information about the 29 movies produced by Marvel Studios.

The data is classified as it follows:

- movie: title from the movie.
- released_date: date at which the movie was released.
- gross_us_canada: cumulative earnings of the movie in US an Canada, that will be our reference value.
- gross_other_territories: cumulative earnings of the mov ie in other territories.
- gross_gross_worldwide: cumulative earnings of the mo vie worldwide.
- alltime_rank_us_canada = all-time gross rank in US and Canada.
- alltime_rank_worldwide = all-time gross rank worldwide.
- budget million: budget of the movie in million dolars.
- rottentomatoes: rating from this web.
- metacritic: rating from this metacritic.com that will be o ur reference value.
- cinemascore: rating from this web.

```
marvel_movies = pd.read_csv("https://raw.githubusercontent.com/JuanArchidona/practica_pandas/main/marvel-cinematic_universe.csv", sep = ",")
```

Now our goal is to process and retrieve the records, with the following necessary information from all Marvel Studios movies:

Tittle, Year, Meta Score, Gross and a field named War_side, that qualified them as a Marvel movie. This parameter will

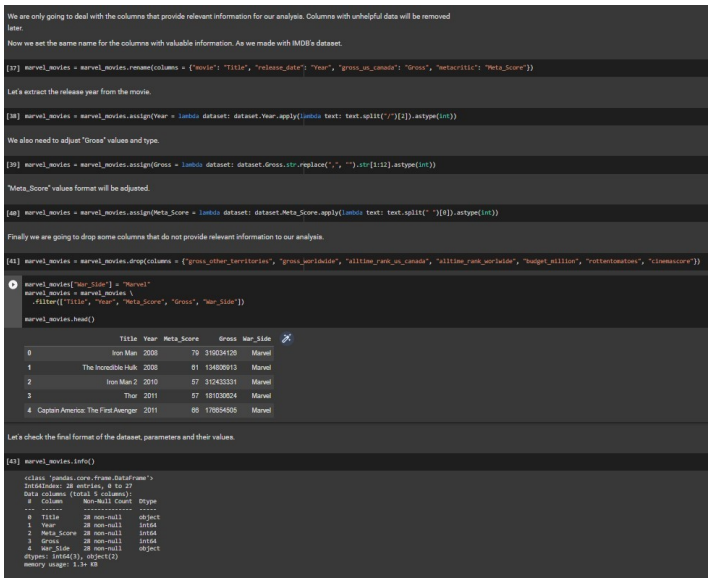
facilitate a later comparison

We proceed to retrieve a record and the data types of each field.

```
marvel_movies.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 28 entries, 0 to 27
Data columns (total 5 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Title        28 non-null     object
 1   Year         28 non-null     int64
 2   Meta_Score   28 non-null     int64
 3   Gross        28 non-null     int64
 4   War_Side     28 non-null     object
dtypes: int64(3), object(2)
memory usage: 1.3+ KB
```

We also can see here that there are several fields that do not provide relevant information to our analysis. Or whose data type does not match the required one.

We proceed to adjust these fields and their values, using methods such as rename(), replace(), assign() or drop()... This is a summary of the steps necessary to make the changes:



We need to recover the best films directed by Martin Scorsese.

Our goal is to adjust their fields in the same way we did with the Marvel Studios movies, to facilitate later

comparisons. Including the War_Side parameter, which also classifies these movies as directed by Martin Scorsese.

```
scorsese_movies = movies.query("Director == 'Martin Scorsese'", engine = "python")  \
  .assign(War_Side = lambda dataset: dataset.Director.str.replace("Martin Scorsese", "Scorsese")) \
  .drop(columns = {"Genre", "Rating", "Director", "Votes", "Marvel"}) \

scorsese_movies.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10 entries, 15 to 836
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Title       10 non-null     object
 1   Year        10 non-null     int64 
 2   Meta_Score  10 non-null     int64 
 3   Gross       10 non-null     int64 
 4   War_Side    10 non-null     object
dtypes: int64(3), object(2)
memory usage: 480.0+ bytes
```

We now proceed to merge both datasets and obtain a sample of the resulting records.

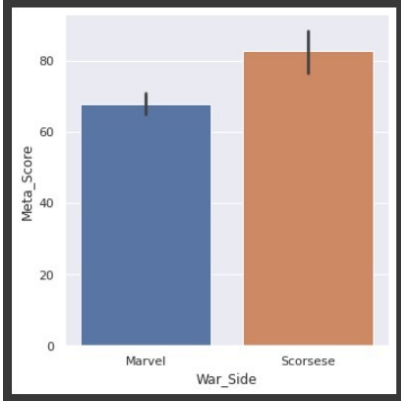
```
movies_war = pd.merge(marvel_movies, scorsese_movies, on = ["Title", "Year", "Meta_Score", "Gross", "War_Side"], how = "outer")
movies_war.sample(5)
```

	Title	Year	Meta_Score	Gross	War_Side
9	Guardians of the Galaxy	2014	76	333718600	Marvel
34	Raging Bull	1980	89	23383987	Scorsese
37	After Hours	1985	90	10600000	Scorsese
32	The Wolf of Wall Street	2013	75	116900694	Scorsese
15	Spider-Man: Homecoming	2017	73	334201140	Marvel

Let the war begin!

First, let's compare Scorsese's and Marvel's films according to their Meta Score, which is a compendium of ratings from the metacritic website.

```
sns.catplot(x = "War_Side", y = "Meta_Score", data = movies_war, kind= "bar");
```




```
movies_war.sort_values("Meta_Score", ascending = False).head(10)
```

	Title	Year	Meta_Score	Gross	War_Side
35	The Irishman	2019	94	7000000	Scorsese
30	Taxi Driver	1976	94	28262574	Scorsese
37	After Hours	1985	90	10600000	Scorsese
28	Goodfellas	1990	90	46836394	Scorsese
34	Raging Bull	1980	89	23383987	Scorsese
17	Black Panther	2018	88	700426566	Marvel
29	The Departed	2006	85	132384315	Scorsese
0	Iron Man	2008	79	319034126	Marvel
21	Avengers: Endgame	2019	78	858373000	Marvel
9	Guardians of the Galaxy	2014	76	333718600	Marvel

Scorsese seems to have the lead in terms of ratings. Now let's compare the films according to their grosses.

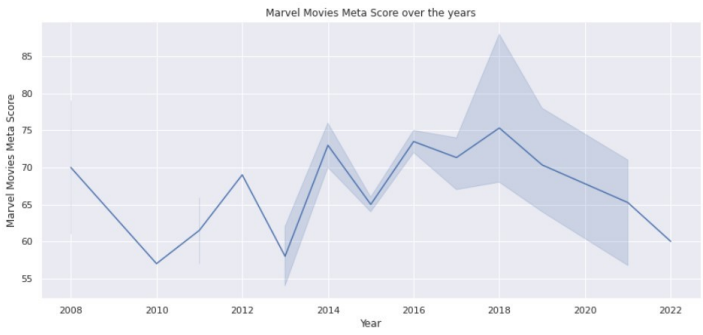
```
movies_war.sort_values("Gross", ascending = False).head(10)
```

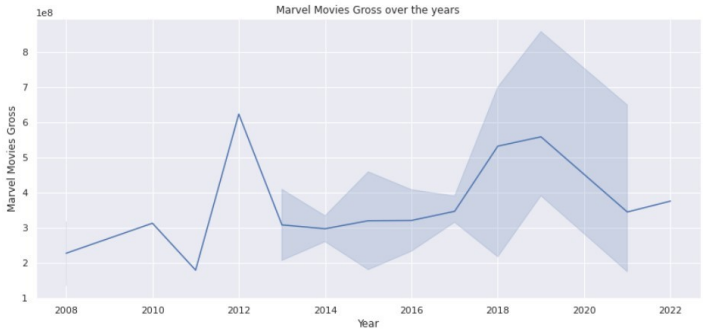
	Title	Year	Meta_Score	Gross	War_Side
21	Avengers: Endgame	2019	78	858373000	Marvel
26	Spider-Man: No Way Home	2021	71	804747988	Marvel
17	Black Panther	2018	88	700426566	Marvel
18	Avengers: Infinity War	2018	68	678815482	Marvel
5	Marvel's The Avengers	2012	69	623357910	Marvel
10	Avengers: Age of Ultron	2015	66	459005868	Marvel
20	Captain Marvel	2019	64	426829839	Marvel
6	Iron Man 3	2013	62	409013994	Marvel
12	Captain America: Civil War	2016	75	408084349	Marvel
22	Spider-Man: Far From Home	2019	69	390532085	Marvel

This battle turns out to be a clear and expected victory for Marvel Studios.

Now let's check if Scorsese's statements made in 2019, may have created groundswell of opinion in the becoming of Marvel movies Meta Scores and Gross.

Clearly it appears that Scorsese's statements made in 2019, could have had some sort of influence on Marvel Studios' movie profits and ratings.





Conclusions

After cross-comparisons of IMDB's top 1000 movies according to ratings, profits, main genre and votes.

To which to add the direct comparison between the films of both sides of the war. And a more than possible influence of the opinion of the acclaimed director.

In my humble opinion there is a clear winner of this contest and he is not wearing a cape: Martin Scorsese.

Publicado por

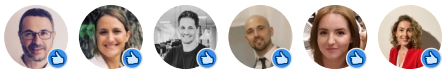
Juan Archidona Ahijado
Data Scientist
Fecha de publicación: 6 meses

1 artículo

Recomendar Comentar Compartir

Daniel Villanueva Jiménez y 5 personas más 3 comentarios

Reacciones



3 comentarios

Más relevantes

Añadir un comentario...

Daniel Villanueva Jiménez • 1er
Data Architect & Lecturer
6 meses

Felicidades Juan. Muy buen artículo!

Recomendar Responder · 1 respuesta

Juan Archidona Ahijado • Tú
Data Scientist
6 meses

Muchas gracias Daniel, de verdad.

Recomendar Responder

Juan José Martínez Quesada • 1er
Data Analyst
6 meses

Gran trabajo Juan !! 👏👏👏

Recomendar Responder

Juan Archidona Ahijado
Data Scientist