Academic Year 2021/2022

Luca Ballarati 867488
Francesco Gregori 889206
Francesco Oliviero 812292

Index

1. Introduction

In the film industry, the sci-fi world of superheroes is dominated by characters from the Marvel and DC universes. Every year films are produced narrating the stories of the protagonists of the comics created by these two great companies.

Marvel Comics was founded in 1939; the Marvel denomination dates back to 1961, when the series dedicated to superheroes created by Stan Lee, Jack Kirby, Steve Ditko and others debut.
The most famous characters are Iron man, Captain America, Hulk, Thor, the X-men and Spider-man.
Marvel Comics is one of the many divisions of Marvel Entertainment, the large entertainment company, founded later in 1998.

The division that deals with film production is called Marvel Cinematic Universe.

DC comics was born in 1934. Among its symbols there are characters who made the history of the superhero and adventurous genre, such as Superman, Batman, Flash, Green Lantern, Wonder Woman.
DC comics is the publishing division of DC Entertainment. The films narrating the deeds told in the comics are produced by the division called DC Extended Universe.

Marvel and DC have always been in competition. There are key superheroes on both sides. Among the audience there are those who adore DC, its characters and the films that have been dedicated to them. And there are those who, on the contrary, love Marvel superheroes, considering the DC universe only a legacy of the past. Which raises the following question: which of the two companies is the better?

This report analyses the film divisions of both companies in order to answer the previous question.

More precisely, the revenues obtained from the films of the two different universes are studied, comparing budgets and receipts, in the opening weekend and totals, recorded in the USA and in the world.
An analysis of the opinions of critics and the public is also carried out through the rating attributed to each film produced.

At first, the data were collected and downloaded, after which a careful analysis of these and of the context was carried out in order to identify those variables that allowed to make an interesting comparison between the two great giants of the superhero science fiction world.
Subsequently, data visualizations were created that allowed the reader to identify the best company from the point of view of the earnings obtained by each film and per year, taking into consideration production budgets and reviews.
Finally, a heuristic evaluation was carried out to highlight any problems with the views, a user test that allowed to evaluate effectiveness and efficiency and a questionnaire to evaluate their satisfaction. At this link you can view the presentation of the project.

2 Design

The dataset used for this analysis was downloaded from Kaggle, an online community of data scientists and machine learning professionals that allows users to find and publish datasets, explore and build models in a web-based data science environment.

The database in question collects data regarding films that have been produced by Marvel Cinematic Universe and DC Extended Universe from 2004 to 2019, the year before the outbreak of the global Covid-19 pandemic. The numbers are therefore not influenced by the latter.

There are 39 films, 23 produced by Marvel, 16 produced by DC. The dataset consists of the following variables:

- **id**
- **Title of the film**
- **Company** = the manufacturer, Marvel or DC
- **Rate** = score from 1 to 10 assigned by IMDb registered users, the most popular source of information about movies and TV series. The user can assign only one vote per film, which can be updated over time
- **Metascore** = score assigned by metacritic.com. It is a weighted average of the reviews of the most important critics and publications. the rating of each review is converted to a scale from 0 to 100 (if the rating is not present it is assigned based on the impression given by the review). The weighted average weights reflect the prestige and respect achieved by critics, and the length of the review.
- **Minutes** = length of the film
- **Year of publication**
- **Budget** = expenditure for the production of the film, in $
- **Opening Weekend USA** = proceeds obtained on the opening weekend of the film, in the United States, in $
- **US** takings = box office takings in the United States, in $
- **Collections worldwide** = box office receipts in all countries of the world, in $

3 Realization of the infographic

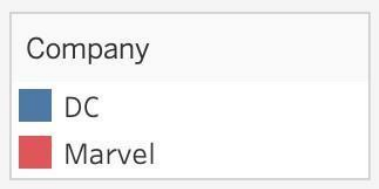
After defining the objectives of the project, it was established which was the most suitable type of graph for visualizing the data. We have chosen to use the scatter plot, a type of graph in which two variables of a data set are reported on a Cartesian space, because it allows to visualize the correlation between two variables and at the same time the distribution of the data.
The data is displayed through a collection of points each with a position on the horizontal X axis determined by one variable and on the vertical Y axis determined by the other.

Subsequently, Tableau was chosen as it is the most suitable tool for creating interactive and professional infographics. Previously collected data was then fed into Tableau and several scatter plots were created to answer the research questions.

3.1 Global vs critical collection

The first question on which we wanted to focus the research was that relating to the relationship between the total income (worldwide) recorded by each film and the judgment of the critics. How are these two variables related to each other?
First, the axes were assigned to the reference variables. The X abscissa axis for the metascore and the Y ordinate axis for the Gross Worldwide.

To distinguish the films produced by the two different companies, the respective colours that distinguish them were used. The colour red for the Marvel Cinematic company Universe and blue for the DC Extended Universe.
The colours are visible in the legend, shown on the right of the graph.

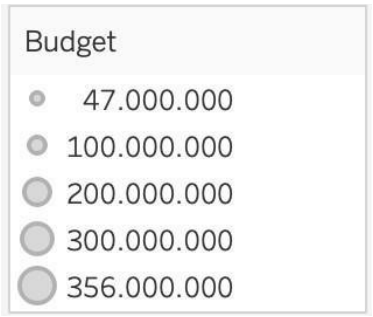


As regards the Metascore attribute, the median with a 95% confidence interval has been added, as the displayed values represent a sample of all possible people's judgments, and not the totality of the population. The choice of the median was made to have a more robust indicator in the presence of outliers.

Differently for the Gross Worldwide the median was used with the interquartile range because the data collected represent the entire population.

For a better understanding of this infographic, some “visual cues” have been inserted: further aspects, necessary to be introduced, because the variables taken as a reference alone would not be able to explain them.

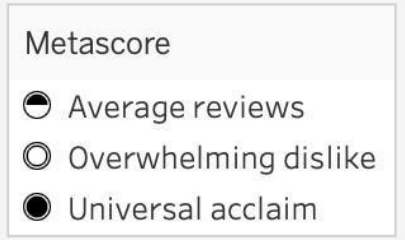
In this phase the dimensions of the markers have been modified following, in proportion, the value of the budget. A higher budget corresponds to a larger dimension of the marker to which it refers.



For the data expressing the judgment it was decided to use the Metascore variable instead of the Rate variable, as the former was considered more authoritative than the latter. Authoritative since it collects a weighted average of evaluations expressed by critics and experts in the sector.
Metascore: the official reference table is shown

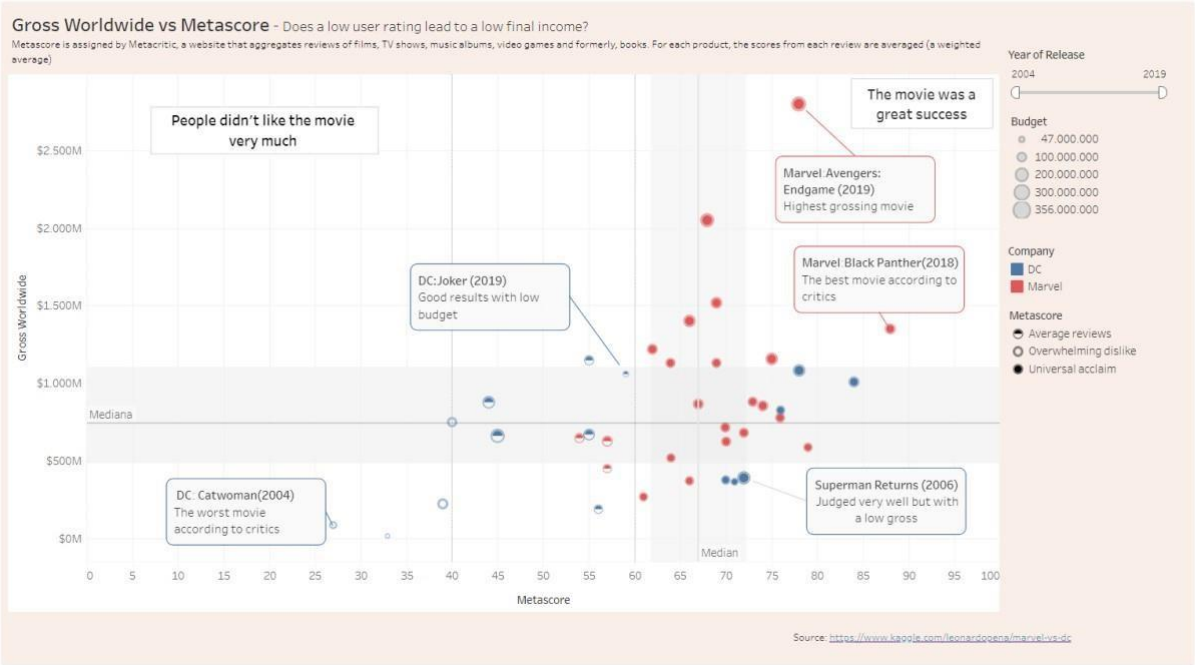
Indication	Video games	Films/television/music
Universal acclaim	90–100	81–100
Generally favorable reviews	75–89	61–80
Mixed or average reviews	50–74	40–60
Generally unfavorable reviews	20–49	20–39
Overwhelming dislike	0–19	

To distinguish the judgments, on the other hand, it was decided to adopt markers following the reference scale for film criticism. In particular, the empty dot for totally negative judgments [0-39]. The semi-full one for judgments that are within the average
[40-60] and the full one for the most successful films [61-100].



A filter on movie release years has also been added to include the time dimension.

As can be seen from the graph, a high judgment of the critics does not always correspond to a high income. It's the example of DC's Superman Returns movie and DC's Joker movie. Furthermore, for these two cases it is also noted that with a lower budget a greater gain was obtained.
However, it can be established that there is a certain dependence between the two variables under consideration.



3.2 Box Office Collections

Through the following dispersion chart, or scatter plot, we wanted to proceed with the analysis of the total income obtained worldwide from each film produced by the two companies and the relative budget used for the production of the latter. There are many ways in which the cinematic success between the two comic book publishers can be compared. Box office receipts are certainly an important figure. It is useful to look, not only at the revenue, but also at the related costs.

The two variables described above were represented on the axes to verify the existence of a relationship or correlation, more precisely to communicate how much the budget of a film affects the respective income.
From the infographic it is possible to identify how there is a positive correlation between total receipts and budget, which means that the values increase together. Interestingly, of all the Marvel and DC films, the highest grossing were the four Avengers (Marvel Universe) films. Consequence of this were also the high budgets used: **Avengers: Endgame (2019)** and **Avengers: Infinity War (2018)** were also the most expensive films.
On the other hand, the film cost less, it is also the film that earned less. We are talking about **Jonah Hex (2010)**, from DC.

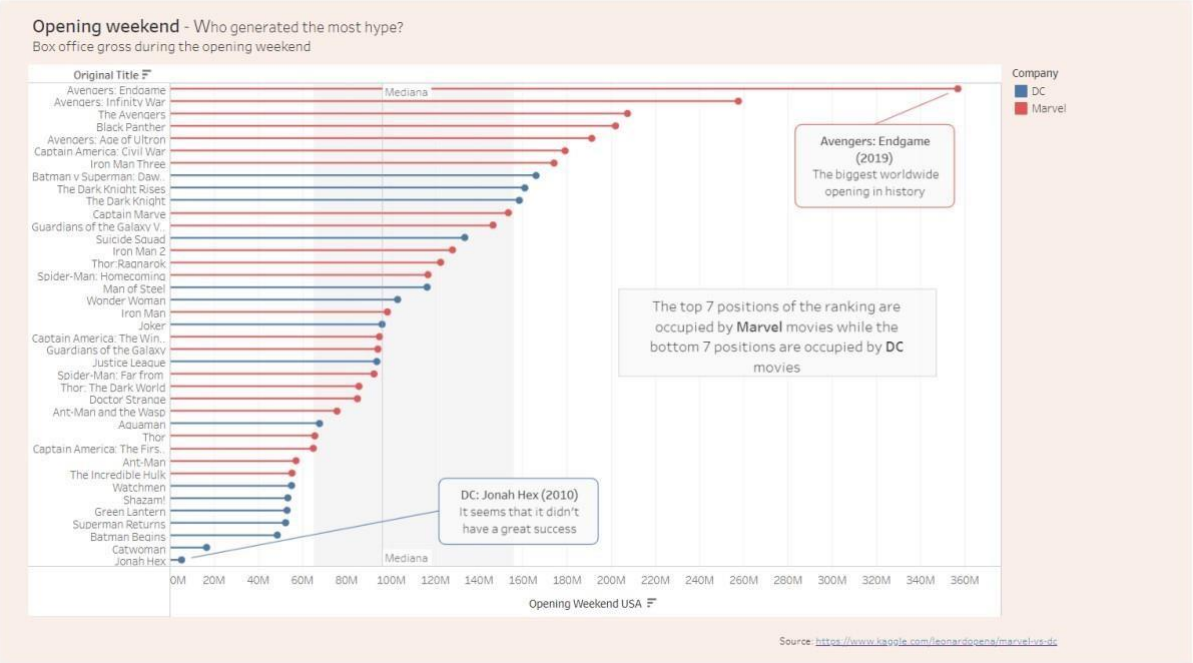
There are some anomalous or extreme values to underline. Honourable mention for **Joker (DC, 2019)** which, despite having one of the most limited budgets available, represents one of the highest grossing for the film division of DC Entertainment. The most disappointing film in terms of revenues, despite a high initial investment, is still

from DC. This is the film **Justice League (2017)**, which cost almost as much as the competition **Avengers: Infinity War (2018)**, but with a much lower income.

The medians and the interquartile range were calculated for the two different variables. The medians divide the display into four different quadrants that highlight the characteristics of the values represented.
The interquartile ranges, on the other hand, allow to estimate the dispersion of the data of the two variables around the median. It can be seen that most of the values are part of the heart of the distributions.

3.3 Opening week

The proceeds during the opening week gives an idea of what expectation the film has conveyed to people. Therefore, a high initial cash out can generally be the result of an excellent promotional campaign aimed at engaging a large number of people.
For this visualization it was therefore decided to use a lollipop chart. With this graph, thanks to a simple descending order, you immediately have a glance on which was the film with the highest initial income and, indirectly, an indication of what the audience expected for this film was. A similar argument can be made in the lower part of the ranking for the film with the lowest grossing.
It is noted that the top positions are occupied entirely by Marvel-branded films, while the bottom positions are all sadly occupied by DC-produced films.



3.4 Profit Analysis

In the infographic below we wanted to analyse the trend of profit over time, trying to obtain information on which films have achieved better performances and which ones have instead led to a loss in economic terms. The profit was calculated with the ROI formula:

$$\left(\frac{(Total\ Revenue - Total\ Cost)}{Total\ Cost}\right) \cdot 100$$

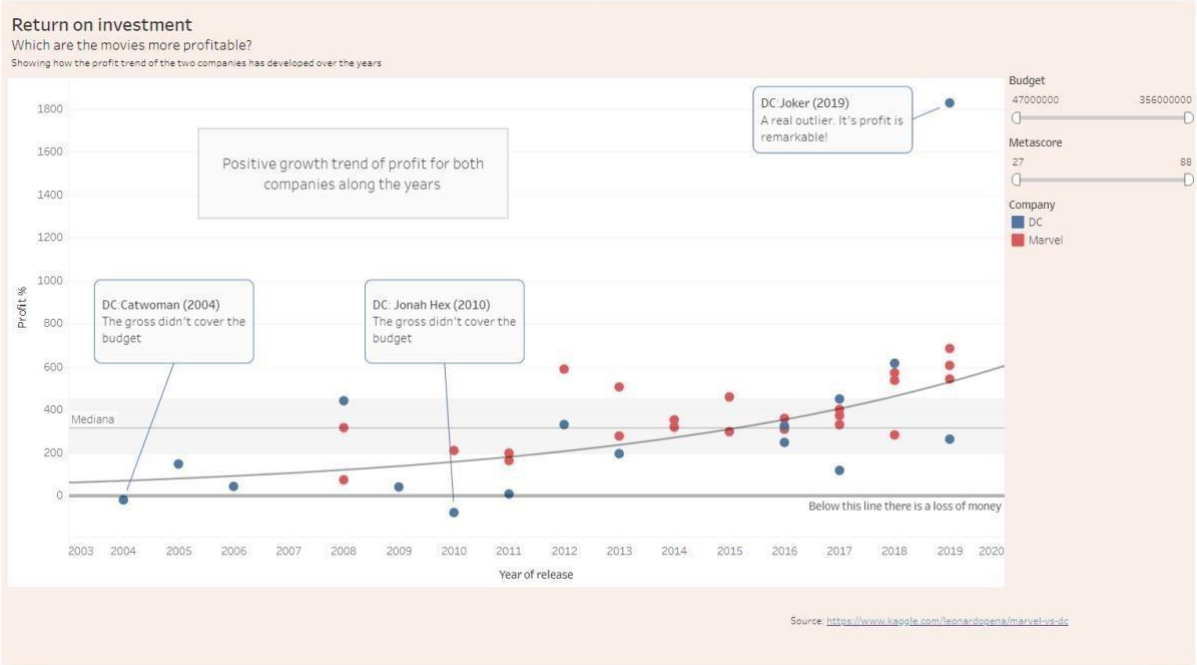
In our case

Total Revenue = Gross Worldwide

Total Cost = Budget

Also, in this case the scatterplot turns out to be a suitable choice for the purpose. Adding colour to the points depending on the film company allows for a third level of analysis, making the distinction between Marvel and DC visible to the user. A regression line was also displayed indicating the profit growth trend over the years. The median line with the band indicating the interquartile range respectively indicate where the median value of the profit values is positioned and where 50% of the observations of the latter are collected.

You immediately notice that there is an outlier (Joker, produced by DC) with a high profit compared to all the others. This film is therefore characterized by having a low budget and a high collection in reference to the budget itself.
A line was placed on the zero value of the profit in order to highlight the two films that reported an economic loss, both produced by DC. It is also possible to apply a filter on the budget and on the Metascore values to narrow the field of view.



4 Quality assessment

A qualitative-quantitative assessment is carried out, consisting of the following methodologies:

- **heuristic evaluation**: process during which external users are asked to interact with the infographic, even thinking aloud, in order to identify any usability problems. The output of this process is a list of usability problems;
- **psychometric** questionnaire: structured questionnaire carried out with the aim of collecting subjective opinions from users;
- **user test**: used to 'make measurements'; it is sufficient to evaluate the *effectiveness*, that is the number of correct answers given by the subjects on the basis of the infographic, and the *efficiency*, which is reduced to measuring the time necessary for people to respond.

4.1 Heuristic evaluation

For this evaluation, three people are involved who are asked to freely use the infographic and comment aloud on the actions that are performed.
The result of this experience has been collected in the following table:

User 1	It is not clear what ROI means. Difficulty understanding what the years refer to in the ROI trend view
User 2	For this user too, the concept of ROI is a bit difficult. The line that divides profit between positive and negative is unclear. The filters on the right help to understand the meaning of the colours, shapes and sizes of the markers
User 3	It has been observed that it might be useful to add a filter on the year of publication in the display of profit distributed over the years. For the rest they are not States found details problems

4.1.1 Corrections

Based on the results of the heuristic evaluation, the necessary changes were made. In the view relating to the trend of ROI over the years, the term ROI has been replaced with a term that is easier to understand, especially for people who are not experts in the economic field. The subtitle of the infographic has also been changed to make the content easier to understand.
Subsequently the title was added on the axis containing the years of publication of the film. Also changed the name of the filter referring to the year of release in all the tabs. Accepting the comment from the third user, a year of release filter has been added to the profit infographic.

Evidently it can be concluded that heuristic evaluation assumes considerable importance in order to improve the usability and clarity of the various infographics.

4.2 User test

Six people were selected to administer the test. After a brief overview of the whole project, they were asked to perform the following tasks:

- Task 1: which film generated the most expectation and which was received with the most indifference. Which companies do they belong to ?

- Task 2: for which films was there a financial loss?
- Task 3: Does a high budget always lead to a high final income? Identify an ' anomalous ' case

4.2.1 Results

In some cases, users were guided through the test. The results have been collected in the tables below.

Task execution times:

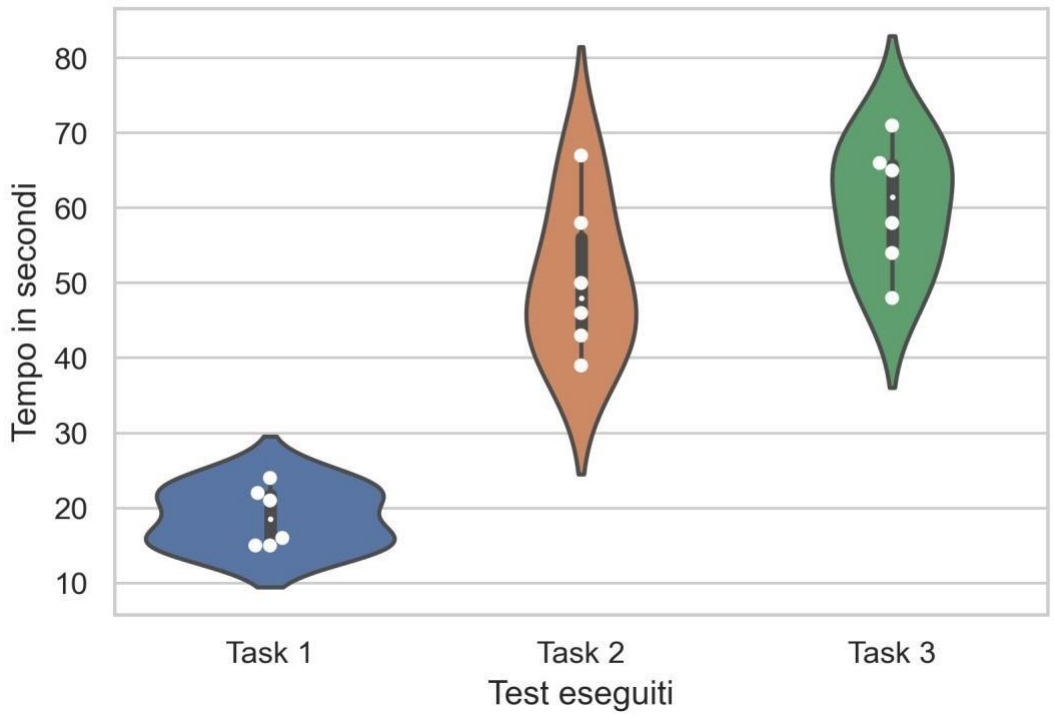
	Task 1	Task 2	Task 3
User1	16	39	65
User2	15	50	48
User3	24	46	54
User4	22	67	71
User5	15	43	66
User6	21	58	58

Mistakes found in tests:

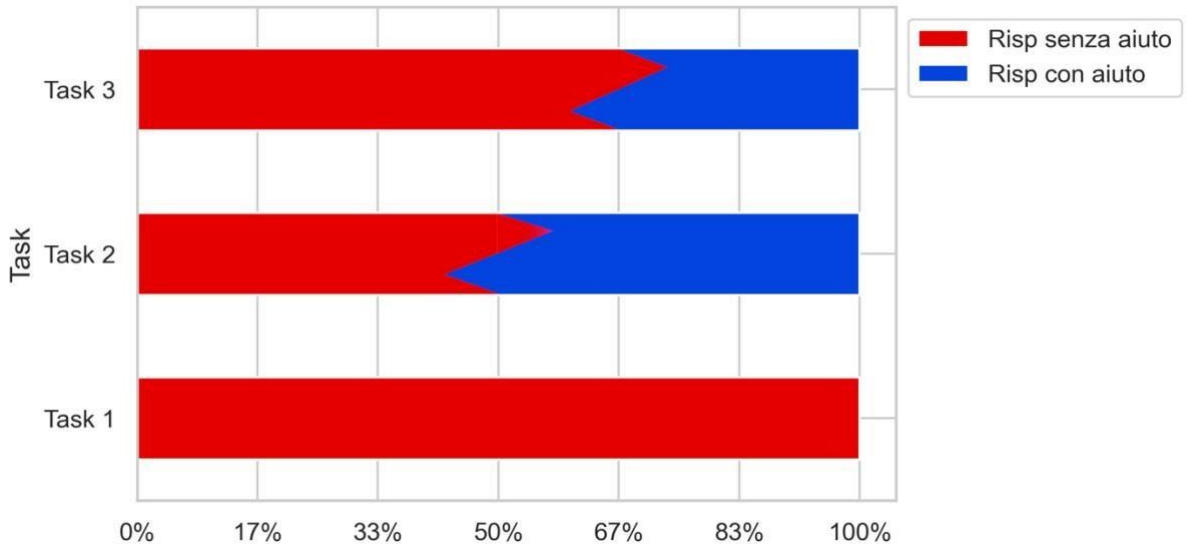
- 0: answer given independently
- 1: answer given with our help

	Task 1	Task 2	Task 3
User1	0	1	0
User2	0	1	1
User3	0	0	1
User4	0	1	0
User5	0	0	1
User6	0	0	1

The result of the time taken to execute the tasks is displayed in the violin plot below



The violin plot allows to visualize the distribution of the values. The overlapping scatter plot shows all the points present. It is noted that the first task was carried out much faster than the others and with very close times between them. The result was expected, since the task actually required was very simple. The outcome of the responses is shown in the following visualization, using a stacked bar plot.



In this case it is noted that the first task was carried out correctly without help, while for the other two it was necessary to provide some more information. The confidence interval is also displayed via the broken line that delimits the two bars.

$$p \pm z\sqrt{\frac{p(1-p)}{n}}$$

4.3 Psychometric questionnaire

Locoro scale [1] was adopted for the realization of the questionnaire. The questionnaire is divided into two sections. In the first, the user is asked to evaluate the quality of the infographic by expressing a judgment ranging from very little (value 1) to very much (value 6). The evaluation parameters are:

- Useful
- Clear
- Disclosure
- Pretty

In the second section you are asked to express an opinion on the 'overall value', also here on a scale of 6 values, with very low as value 1 and very high as value 6.

The questionnaire was administered to 12 people and the collection of opinions was carried out through a Google form.

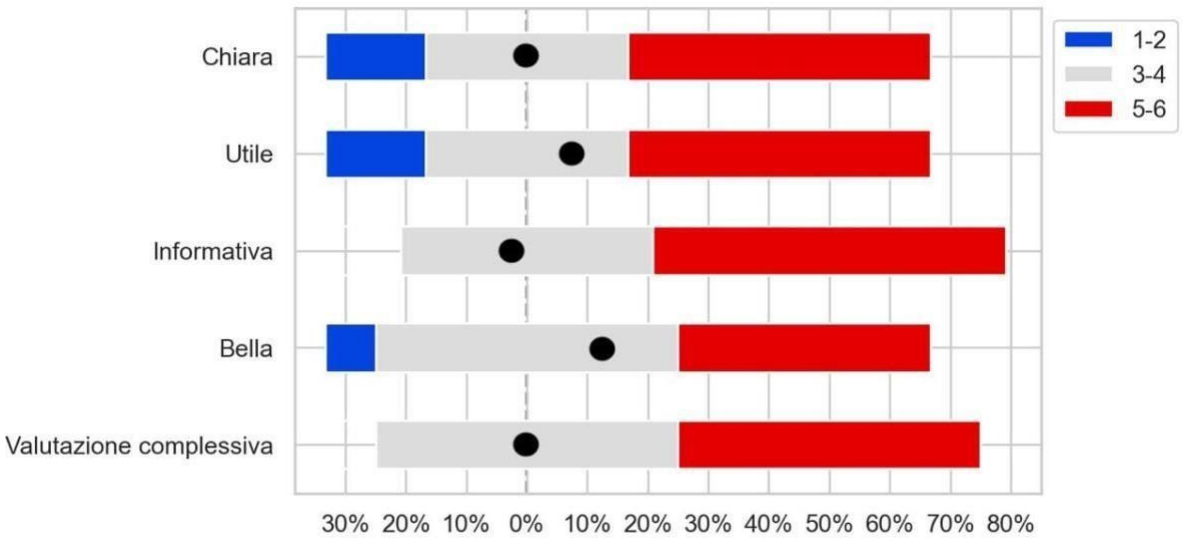
4.3.1 Results

The following table contains the outcome of the questionnaire with the sum of the various values for each evaluation

Classes	Clear	Useful	Disclosure	Pretty	Total value or
1	0	1	0	0	0
2	2	1	0	1	0
3	2	1	3	1	3
4	2	3	2	5	3
5	5	5	5	4	4
6	1	1	2	1	2

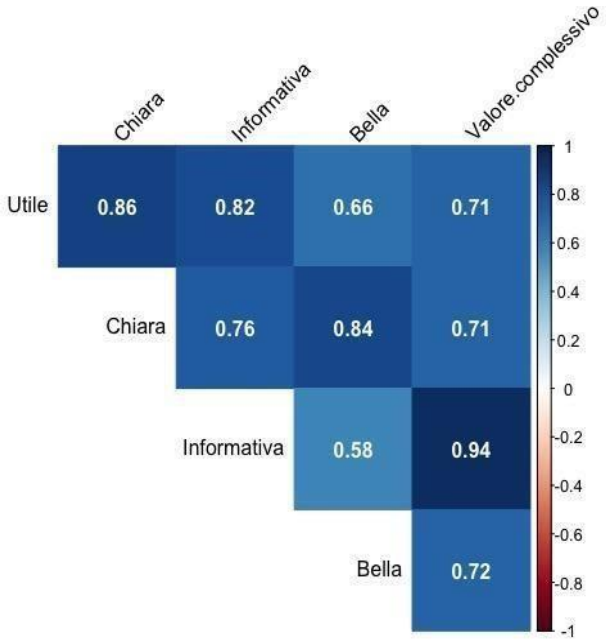
The information contained in the table was then analysed through two types of views: diverging stacked bar chart (or butterfly chart) and the correlogram

- Diverging stacked bar chart



It can be seen immediately at a glance that the evaluation is overall positive, in particular with regard to the information aspect and the overall evaluation.
Within the central values it was highlighted, by means of the black dot, whether the judgment is more shifted towards 3 or 4, considering the average value of the relative evaluations.

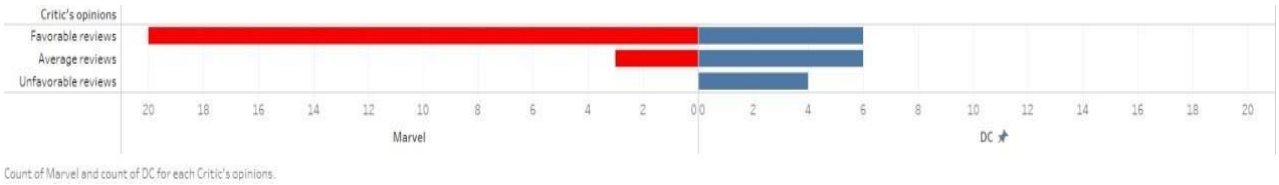
- Correlogram



The graph allows to show the correlation between the different variables involved. The colour is used to give an immediate glance on extreme cases, i.e., those with higher correlation and those with lower correlation.
There is a strong positive correlation between the informative aspect and the overall value of the entire visualization, while the lowest correlation is found between the beautiful and the informative aspect of the same.

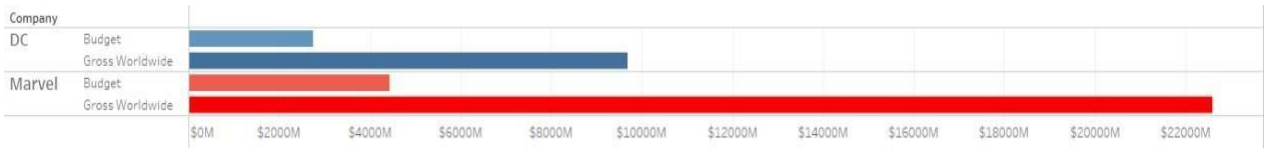
6 The final winner is ...

From the graphs below it can be concluded that as regards the aspects of review and collections, the winner is Marvel.



Count of Marvel and count of DC for each Critic's opinions.

The average metascore achieved by Marvel movies is 68.65. 12 points higher than the average metascore obtained from DC production films, equal to 56.50. The graph above shows that most of the Marvel films are judged very well by the critics (20 films with reviews in favour). Notable is the lack of films with negative reviews. As for the rival company, the films are evenly distributed across the 3 review classes.



Budget and Gross Worldwide for each Company. Color shows details about Budget, Gross Worldwide and Company.

A second fact that confirms the predominance of Marvel Cinematic Universe on DC Extended Universe is the total earnings that the two companies have achieved in the time frame analysed.

Marvel's earnings for the production of the films in question are $ 18,151M, nearly 3 times higher than DC's earnings of $ 6,944M.

5 Conclusions and future developments

It would be interesting to deepen the research question to understand in detail if there may be further variables to consider in order to reach more accurate conclusions. For example, consider movies that were not present in the source dataset. Among these we can mention for DC the X-Men saga: X-Men Final Conflict (2006), *Wolverine The immortal* (2 013), *X - Men Days of a future past* (2 014), *X - Men* The *beginning (* 2011), *X - Men the origins Wolverin and* (2 009) and for Marvel: The Fantastic i 4 (2005), Fantastic 4 & Silver Surfer (2007), Fantastic 4 Fantastic Four (2015).

Another consideration that can be made is to be able to add information relating to the prizes awarded and the nominations received, with a particular reference to the Oscars.

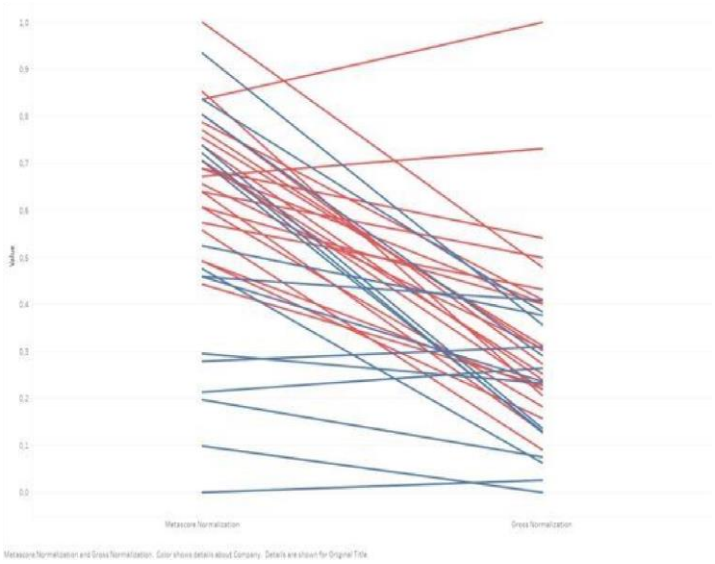
An area of interest could be to investigate the geographical distribution of receipts to identify possible user clusters.

More recent data could be collected including information regarding the quality of films during the pandemic, to be correlated with reviews of films released in earlier periods.

What characteristics must a superhero film have to be considered successful?
In general, a high budget leads to high receipts and valuations, but anomalous situations can still occur. For this reason, it can be said that a good initial investment tends to be a fundamental aspect for a successful film, but it cannot be the only variable to be taken into consideration.

It could be interesting to understand if there are data available to evaluate whether it is the criticism that contributes to the success of the film or whether its contribution is irrelevant. From the parallel plot below, it can be seen that in some cases a high rating leads to lower takings compared to other films with lower reviews.
The scales have been normalized in order to have the values within the range [0,1]. The normalization formula used is the following:

$$Z = \frac{x - \min(x)}{\max(x) - \min(x)}$$



7 References

[1] Actis-Grosso R. & Batini C. Locoro A. Cabitza F. "Static and interactive infographics in daily tasks: A value-in-use and quality of interaction user study." In: Computers in Human Behavior (2017).

Link to the dataset used:
https://www.kaggle.com/leonardopena/marvel-vs-dc

Link to the software used for the visualizations:
https://www.tableau.com

Link for consultation on data visualization: https: //www.data-to-viz.com