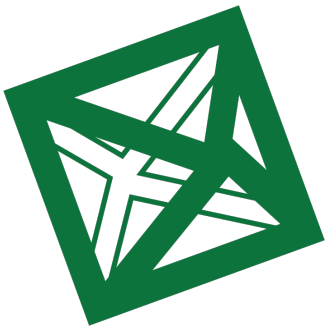
UNIVERSITÀ DEGLI STUDI DI MILANO
BICOCCA

DEPARTMENT OF INFORMATICS, SYSTEMS AND COMMUNICATION



Master Degree in Data Science

Data Management Project

# Evaluating the relationship between games and reviews

PAOLO CAGGIANO - DAVIDE GIARDINI

# Indice

# 1 Introduction

The world of videogames is constantly growing. According to game's market research firm Newzoo's report, already by 2021 the industry was well near to the 3 billion users mark (figure 1). Furthermore, the industry is still on an impressive growth (+5,4% from 2020 to 2021) as it was one of the of the few that was positevely impacted by the COVID19 pandemic.

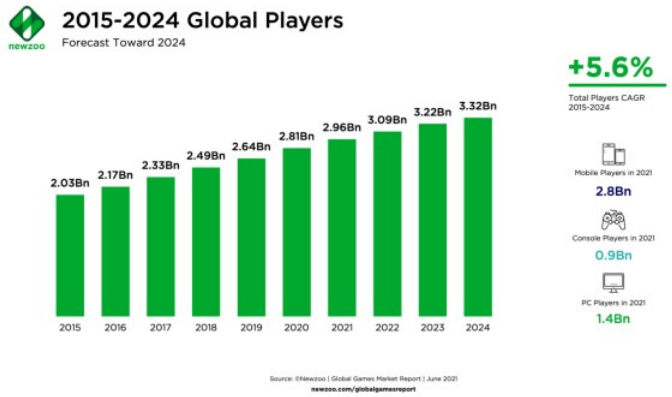



Figure 1: Global players from 2015 to 2021, plus global players prediction through 2024

According to the Entertainment Software Association's (ESA) 2022 report, in the US the average video game player is 33 years old, and 76% of players are over 18 (figure 2). The average time spent playing games each week is 13 hours.

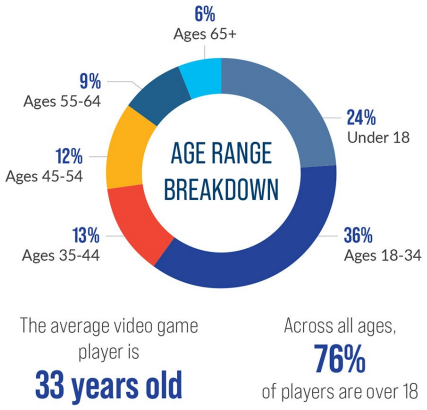



Figure 2: Age breakdown of video games players in the US in 2022.

This huge crowd of enthusiast make the video games market one of the most profitable. According to the NewZoo 2022 report the global market as a whole generated in 2022 almost 200 Billion dollars, with the Asia-Pacific region accounting for almost half of that (figure 3). According to the same report more than half (53%) of the 200Bn$ revenue has been achieved by mobile games alone. The second most profitable platforms are consoles (27%) and then, at the third place, PC games account for 38.1Bn$ (19%).

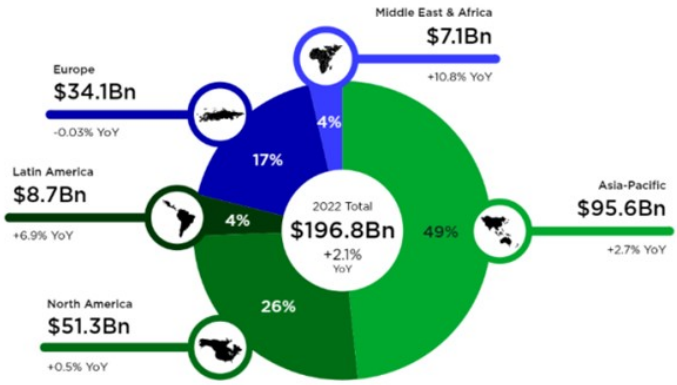



Figure 3: Global games market per region in 2022

But how does this market work? How do you make a good game, both in user's review and profit terms, in an industry with this competition? Furthermore: it is clear that in this type of environment, as it is part of the entertainment market, reviews are extremely important. But how much important? Are professional critic's review the main factor of a game's success? Or, as the game community is strictly bound to the internet, other users' reviews are what count the most? This project was born to address these, as well as other, questions: how much are the reviews important in terms of success or failure of a game? Are there genres more attractive than others in terms of reviews?
To answer this interrogations we need to face multiple problems. First of all, there is not a ready to use dataset where all this information is stored, so we have to build it ourselves. Second, these type of information, like the characteristics of a game, its price and its reviews, are not present together in the same source of data. It is therefore clear that there is the necessity to integrate different sources. Third, the game industry doesn't provide an unique type of game's identification. To achieve this integration, we have to rely on the game's title which, in many cases, can vary in slight ways between different platforms. Hence, it will not be easy to recognize the correspondence and resolve the conflict between the two datasets.

# 2 Data Acquisition

As previously said, there is not any ready-to-use dataset in which all the informations we need are stored. For this reason, we needed to choose multiple sources of data. We decided to focus primarily on PC games, as its primary seller, "Steam", is the perfect source for users' reviews, game prices and genres. It is therefore important to underlined that the results of our research are not universally valid, but they refer to PC games only. For what regards critics' reviews, we used the website "Metacritic". We found it to be the best source of data because it gives every game a score by averaging all others major critic's websites' scores. In this way, our findings wouldn't have been significant only for reviews coming from a specific website, as the scores we used are computed from all the major platforms that reviewed the game.

## 2.1 Steam

Steam is a video game digital distribution service developed by Valve and launched in September 2003 that runs on macOS, Linux and Windows. The Steam client allows users to install PC games online directly to their cloud drives after purchase.
Steam is a giant in the game industry. According to an article published by pcGamesN more than 50 thousands games are available on Steam. In the 2021 Year in Review report by Steam itself, we can also see that Steam was accountable for 38 billion hours of play time in 2021 (that is 4.3 million years). It is also written that there are 132 million monthly active players in the platform and that in 2021 alone it delivered 32.9 exabytes of data.

To extract informations from the site we used a scraper, made in Python thanks to the Selenium and BeautifulSoup libraries.
We firstly scraped the global top sellers page (figure 4) to extract the games our investigation will be based on. After using Selenium to scroll down the entire page, in order to load onto the browser all the 5771 games of the list, we retrieved for every game its title, position on the ranking, release date, price, and steam code.
We then used this last element to open the game's page (figure 5), from which we scraped the number of reviews made by the users, the ratio of positive to negative reviews, and the list of genres that best describe the game. To perform this last action, we needed to tweak Selenium's Chromedriver's cookies allowing the view of violent content, in order to bypass Steam's age check.
Lastly, the attribute "toDrop" was set to True in two cases:

1. The page was significantly different from a common game's page, and therefore the scraping would have resulted in an error. This happened especially for articles that were not games, like the Steam Deck.

2. Sexually explicit content, that we decided to exclude as it falls under a different use and a different userbase than the ones of a typical videogame.

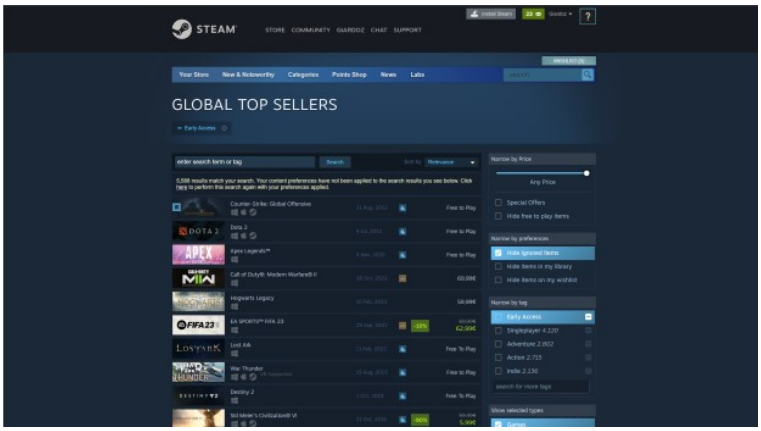



Figure 4: Global Top Seller Steam Page


Figure 5: FIFA 23 game's page on steam

## 2.2 MetaCritic

The second source of data is Metacritic, a website that includes reviews in the fields of music, game, tv and movies. The Metacritic's score, associated to every game, is calculated by computing the average of all the scores the game has achieved in various major game critics' websites.

To scrape all Metacritic's PC games reviews, we connected via Selenium to every Metacritic's game reviews' page (786) (figure 6). At the end, the scraper had downloaded title, Metascore and release date of 78456 games. Of this reviews, though, only 5251 had a Metascore. This is because the score is computed only when 4 others critic's reviews are available. We decided that we wanted whatever review was obtainable so, if a game didn't have a Metacritic score, the scraper would go to the game's page on Metacritic and compute itself the average of whatever critics where on the page (figure 7). In this way, we added

a new review score to 8267 games, achieving a total of 13518 game with a score.

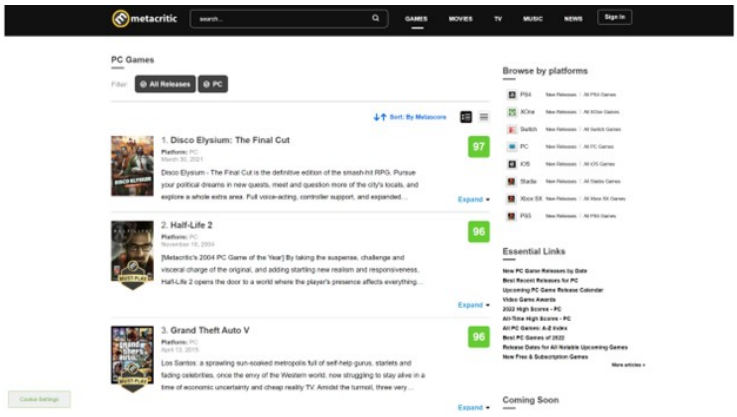



Figure 6: Metacritic's PC games page

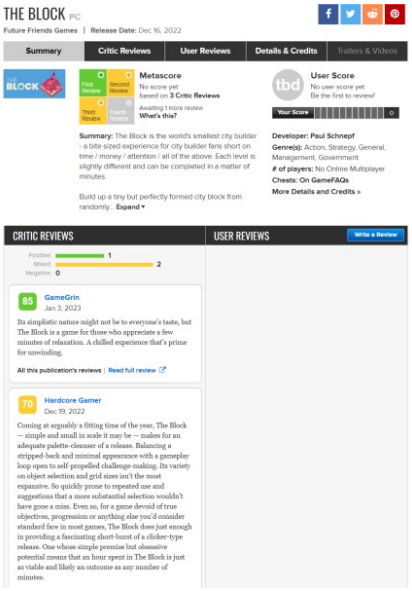



Figure 7: The block's review page. The metacritic score is not available (TBD), but there are 3 critic's reviews from which we can compute the average.

Lastly, we soon noticed that some steam games had no review under the Pc section of Metacritic, but had them under other platforms sections. We therefore decided to scrape games' reviews also in these other sections, and to use an average of the scores of the reviews found as a proxy of the Pc review, for those game where this last was not available.

The same scraping process was therefore applied to all the other sections of Metacritic's game reviews, namely: "ps5", "ps4", "xbox series x", "xbox one", "switch", "ios" and "stadia". We retrieved 26432 games, 20317 of which had either a metacritic score or some other reviewer's score.

# 3 Data Storage

For the storage of the data we used Mongodb. We used a NoSQL database because we didn't know a priori what the schema behind the data retrieved would be. A NoSQL database provided us flexibility in the creation of documents. Furthermore, Steam's videogames can have multiple genres, and this list can always be of different lengths. Once again, it was the flexibility behind this type of database that made us want to use it.
From all the types of NoSQLs, we opted for a document based because we thought that a JSON like format was perfect right from the start. In fact, we have soon noticed that, when scraping, the use of python's lists and dictionaries to momentarily store the data was very convenient. From that point, with the result of the scraping being a list of dictionaries (each representing a game), storing permanently the data in a JSON format was the choice that seemed most obvious to us.
Lastly, we knew that our research would have been strongly game oriented. We knew we would have never queried the review table alone, but always queried the games based on their characteristics and their reviews' scores. For this reason, we thought that the "embedding" approach that a document based database uses would have been more functional (and consequently more efficient) than for example the separate table structure of a relational database, that would have required a join operation for each of our queries.

We used Python's library "Pymongo" to interact with the database. Within our database we built three collections named: Steam, MetacriticPc and MetacriticAll. The last two contained respectively the data scraped from all the Metacritic's reviews from the Pc section, and from all the other sections. After the data integration of this three, we would have gone on building a fourth final collection, that stored the games with their respective reviews' scores embedded.

# 4 Data Integration

## 4.1 Data Cleaning

As we previously stated, the game industry doesn't have any unique serial number we can use to merge the two datasets, so we used the title. This means that we should have done everything possible to normalize the titles in the two datasets, deleting every possible source of misalignment we can find.

For every game in the Steam collection, we proceeded to create a new attribute called "normTitle" in which we stored the title after:

- removing its punctuation
- turning every letter to lowercase
- removing companies name like "ea sports" (eg: "EA SPORTS FIFA 23")
- turn every roman letter into its arabic version

From the same dataset we deleted:

- games in which the attribute "toDrop" was equal to True.
- duplicates
- games in which the release date was "coming soon" or after the time of the cleaning.
- articles with "bundle" inside the title, as they are a collection of different games.

This deletion accounted for 562 games, lowering the number of observation in our dataset to 5209.

We then added another normTitle attribute, normTitle2, that removed from the title also words like:

- deluxe edition
- enhanced edition
- definitive edition
- complete edition
- gold edition
- game of the year edition
- goty edition
- season pass edition

- special edition
- anniversary edition
- expansion
- pack

The merge would be firstly done with the normTitle attribute, so that if it was available a review exactly of that particular edition it would match. In the majority of cases, where there is no review written for a specific edition, the addition of normTitle2 would be used to merge the game with its general review. This was done because all this special editions are often only different from the original by details: the addition of customizable characters, new weapons unlocked, a DLC incorporated with the purchase, ecc. We therefore thought that it would be a fair assumption to use the general game's review's score as a proxy of any of its minor editions. Of course, if the game without the keywords already existed in the dataset, the attribute normTitle2 was not stored, as in the merging phase it would have caused duplicates.
As the last thing, we normalized the release date and the prices. For this last ones, we removed the € symbol and set all "Free To Play" games' price to 0.

For what regards the MetacriticPc collection, we built the normTitle attribute in the same way as above. We normalized the release date and deleted the duplicates. We did the same cleaning to the MetacriticAll collection, after computing the average of the same game's reviews through the different platforms.

## 4.2 Data Merging

In the previous phase of cleaning, we assured to resolve the conflicts that may occur between the two different source. The first (1) and proper merge was done by combining the Steam dataset with the MetacriticPc dataset, by the attribute "normTitle". We then proceded to solve the integrations problem by:

2 Merging the games that had not already found its correspondent review with the MetacriticPc dataset by normTitle2.

3 Merging the remaining games with the MetacriticAll dataset by normTitle.

4 Merging the remaining games with the MetacriticAll dataset by normTitle2.

Summarizing, these are the numbers of correspondences we found for every merge we have done:

1. With the first merge we found 4275 correspondences between the two datasets. Of this 4275 games, 2744 had a review score, while the other had neither a metascore nor other critics' reviews.

2. With the second merge (i.e: on MetacriticPc by normTitle2) we merged 21 new games and we added 31 new games to the ones with a review score.

3. With the third merge, we found a correspondence between 93 new games and we added a review score to 334 games. score.

4. With the fourth merge, 3 new games where connected to their review and 4 new review scores were added to the corresponding games.

In conclusion, of the 5209 games we scraped from Steam, we have been able to find a correspondence with the Metacritic dataset for 4392 of them (84,3%). Of this 4392, 3113 had either a Metacritic score or other critics' reviews' score that we could use for our analysis (71% of the merged games).
The operations we put in place to resolve possible problems in the integration of the two datasets accounted for 117 of the matches and 369 of the games with a score.

After the merge was done we uploaded the resulting dataset to MongoDb, creating a new collection named "GameReviews". This collection contained all the 3113 documents of the games that had a score to work with. In this collection, every document was composed of:

- **id**: Object Id.
- **rank**: position on the Steam Top Selling list.
- **title**: game's title.
- **relDate**: game's release Date.
- **price**: game's price.
- **posPercRec**: percentage of positive recent users' reviews.
- **totRec**: total number of recent users' reviews.
- **posPercTot**: percentage of positive total users' reviews.
- **totTot**: total number of total users' reviews.
- **genres**: Array of genres that define the game.
- **normTitle**: normalized title.
- **scorePc1**: average score of the game between critics' reviews.

# 5 Data Quality

The quality of the data is an important aspect to take into account. Poor data quality implies wrong analysis.
We have analyzed two measures of data quality: *completeness* and *consistency*.

## 5.1 Completeness

For what regards *tuple* completeness, in our final dataset we have six games with null values in: posPercRec, totRec, posPercTot, totTot. In other words they don't have any data regarding Steam users' reviews. For what concernes *object completeness*, of the 5209 games that were available on Steam global top selling page( without considering todrop games we talked earlier), our final dataset now includes 3113 of them (60%).

## 5.2 Consistency

The goal of the consistency analysis was to investigate why 817 games didn't find a correspondence in the Metacritic dataset.
We wanted to understand if these games didn't merge because there was a different representation of the same title in the two different datasets, or just because the game's review was not available in Metacritic. To carry out this analysis, we have extracted a random sample of 10% of the 817 games and manually searched for a correspondence in the Metacritic website (figure 8).

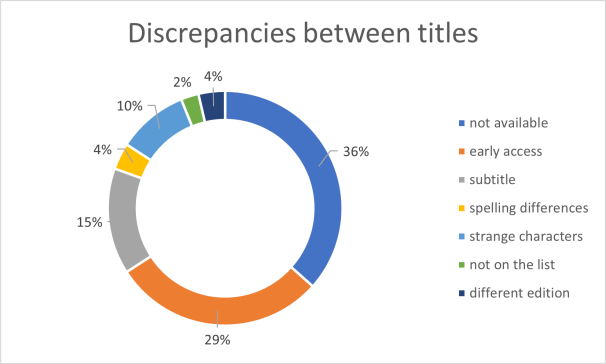



Figure 8: Title consistency analysis

From this analysis we discovered that the major problem is related simply to the absence of the game in the Metacritic website. This accounts to 36 % of the discrepancies between the titles in the two datasets. Another 29 % of the games are in "Early Access". This means that the game was still in development, and therefore didn't appear in the Metacritic's list of games we scraped data from. We figured out that this was not a problem, as critics would wait for the official game release to write a review anyway. Other problems include different ways of writing the title: the presence of subtitles(15%), differences in spelling (4%) and the presence of strange characters(10%). Other minor problems are related to different editions of the game (4%) and to the fact that the game was not available in the list we scraped from for unknown reasons.

# 6 Exploratory Data Analysis

We now proceed to use exploratory analysis and visualizations to address the research's intial questions. The first question we want to address is related to the relation between users' (PosPercTot) and critics' (scorePc1) reviews.

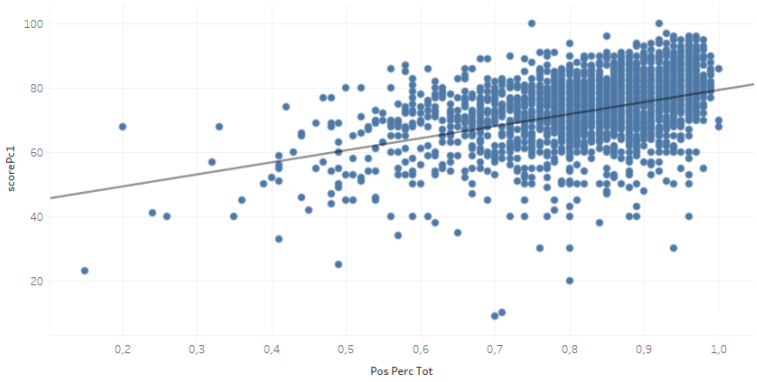



Figure 9: Relationship between users' and critics' reviews

The graph in figure 9 shows that there is an affinity between these two quantities: the users' reviews and the critics' reviews are often in accordance. The correlation, though, is not as marked as someone may expect.

For what regards the prices of the games, the median price is equal to 20€. As the boxplot in figure 10 shows, even if most of the games stand under the 50€ mark, some of them reaches very high prices. "Tales of Arise", "PGA Tour 2K23", "NBA 2K23", "Final Fantasy 7" and "Naruto to Boruto" are the most expensive games present in the steam top selling page, as they all reach 80€ in price.

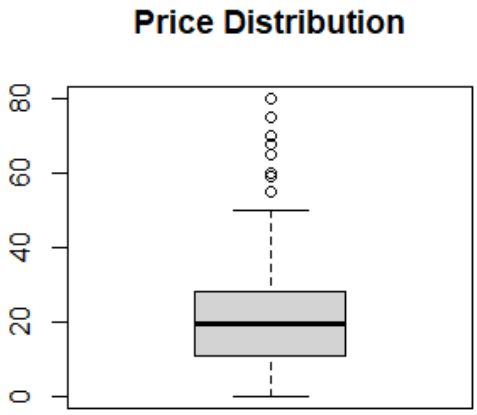



Figure 10: Game's price distribution

It's also interesting to analyze what are the most expensive genres.

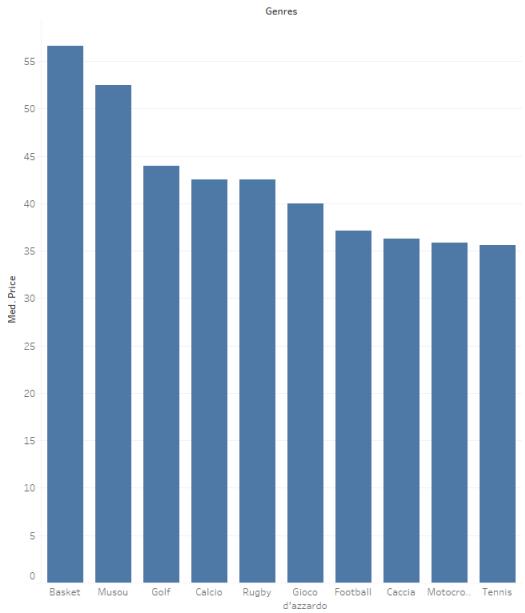



Figure 11: Most expensive genres

The graph in figure 11 shows that the most expensive games, in terms of medium price, regards basketball videogames: the price is about 55 €. In general we understand that sport related videogames are those with an higher average price: Basket, Golf, Football, Rugby, Gambling, American Football, Hunting, Motocross and Tennis are 9 of the 10 most expensive genres.

The next table, instead, shows the 10 genres with the higher medium critics' score. Games in the genres of "360 videos" and "Software" seem to be those that tend to achieve an higher score between critics.

| Genres | |
|---|---|
| Video a 360 gradi | 86,000 |
| Software | 86,000 |
| Skateboard | 84,500 |
| Gioco d'azzardo | 83,250 |
| Sviluppo di giochi | 83,000 |
| Basket | 82,667 |
| Baseball | 82,500 |
| Spietati | 82,400 |
| Immersivi | 82,000 |
| TrackIR | 81,167 |

Figure 12: Medium score per genre

Last, but not least, we investigated how the reviews between critics and users influences the number of downloads. For what concerns this last measure, we decided to use the number of people reviewing the game as a proxy of the number of people that downloaded it. A partial answer is given by the graphic in Figure 13. On the left side the number of reviews are compared with the score of the critics'. On the right with the score of users'.

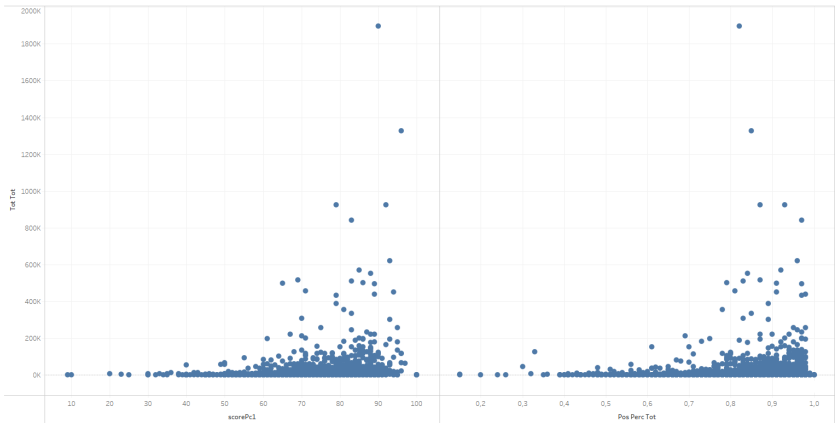



Figure 13: Number of reviews and Reviews' Score

Both the graphs show how the most downloaded games tends to be the ones that received positive reviews between both critics and users. Using "Rank" as a proxy of the number of downloads instead, the correlation doesn't appear strong. There is only a slight decrease in the critic's score when we pass from the games that rank higher (with lower numbers) and games that are at the bottom of the Steam's top selling game page (with higher rank numbers), as shown in the figure 14.

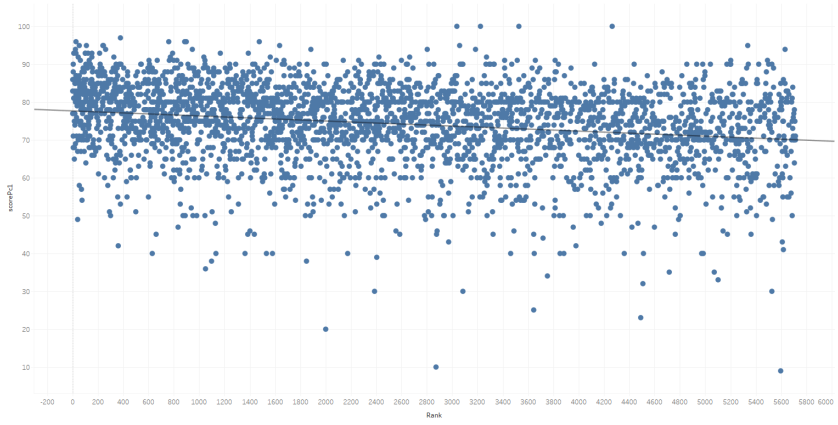



Figure 14: Critic's Reviews' Score and Rank in the Top Selling page

# 7 Conclusion and Future Developments

In conclusion, we have firstly found a discreet agreement between critics and users in term of reviews' score. We have also seen how the most expensive games are related to sports, even if not all of them appear in the top ten critics' favourites. Lastly, we have found a light relationship between both critics' and users' reviews and the number of downloads.

In a future research we could improve our work by finding a new measure for the number of downloads. As we soon noticed in the exploratory analysis, the position of a game inside the steam top selling page(rank) was not a good proxy of the number of downloads. This could be either because the page shows "current" top selling games (but Steam doesn't specify any specific time constraints for "current") or because the page shows only what games sell more than others, but doesn't quantify how much they oversell the next. We tried to use the total number of users' reviews as another proxy, and that gave us more interesting results, but other measures could be taken into account.
We were not able to find any data regarding the true number of downloads of a game on Steam (probably because the platform itself doesn't make them available to the public), but we have found in sites like "steamcharts.com" information about the number of concurrent player in realt time, as well as its 24-hours and all-time peaks. Furthermore, it would be really easy to enrich our dataset with this information, as steamcharts.com uses the same game's serial number steam uses. Maybe, this information could turn out to be a better proxy than the one we used.