2025

Steam Games 2013 to 2023

DAVID NOGUEIRA

Index

1 - Contextualization 2

2 – Proposed Statement 3

3 – Project Description 4

4 – Business Intelligence Questions 5
4.1 – What is the best-selling game over the years (2013-2023)? 5
4.2 – What is the highest-priced game on the platform? 6
4.3 – Which game has the most positive feedback? 6

5 – Business Analytics Questions 7
5.1 – Data Correlation Analysis 7
5.2 – What are the factors that most affect a game's sales? 8
5.3 – According to a set of factors, will the game receive positive feedback? 10

6 – Share 13

7 – Conclusion 14

8 – Bibliography 15

1 - Contextualization

This project was developed as part of the **Data Analyst** course at **CESAE Digital**, with the goal of consolidating knowledge in R through autonomous exploratory and statistical data analysis.
Each student had the freedom to choose the dataset to be analysed, and this project focuses on exploring the "***Steam Games 2013 to 2023***" dataset, obtained from the Kaggle platform.
The original report was written in **European Portuguese**, as was the **R code**, which retains variable names and annotations in the original language. However, for portfolio purposes, the **report** has been **restructured and translated into English**.

2 – Proposed Statement

The statement defined by the instructor for this project aimed to guide the execution of a critical data analysis using R.

In this phase, it was necessary to select a suitable dataset for the analysis. After making this choice, R code was developed for loading and performing a preliminary exploration of the dataset. Additionally, the creation of a report in Word or, alternatively, a PowerPoint presentation was required, including an initial description of the chosen dataset.

As part of the analytical process, two Business Intelligence (BI) questions and two Business Analytics (BA) questions had to be formulated, justifying their relevance to the analysis.

Once these requirements were completed, the next step of the project involved answering the BI and BA questions based on the available data.

3 – Project Description

The R programming project aims to conduct a comprehensive analysis of game data available on the Steam platform from 2013 to 2023. Steam is one of the largest digital distribution platforms for PC games and software, offering a wide variety of titles across different genres.

The analysed data consists of a set of relevant information about the games, including the game title, release date, price, number of positive and negative reviews, category identification (app_id), and the maximum number of users.

The main objective of the project is to explore this data to identify interesting insights into the Steam gaming market over the years.

The project includes data visualization through charts and tables to facilitate interpretation and communication of the results. Additionally, data preprocessing and cleaning techniques (Excel, Power BI, and R) were used to ensure the quality and integrity of the analyses.

Finally, the goal is to provide valuable insights for game and software developers, market analysts, and gaming industry enthusiasts, contributing to a deeper understanding of the gaming landscape on the Steam platform over the years.

4 – Business Intelligence Questions

4.1 – What is the best-selling game over the years (2013-2023)?

This question is essential for understanding the popularity and commercial success of games on the Steam platform over time. Identifying the best-selling game can provide valuable insights into player preferences and market trends. Additionally, this information can be useful for game developers and investors when assessing the profit potential of new projects.

Since my dataset does not have a "sales" column, I analysed the maximum number of owners, as each player had to acquire the game. However, some games are free (price = 0) but still highly profitable due to in-game transactions. Therefore, I divided this question into two analyses: one excluding free games and another including them.

```
> jogo_mais_vendido_sem_gratis
          name release_date price positive negative  app_id max_owners
45369 New World         2021 39.99   154914    73900 1063730  100000000
```

The best-selling game, excluding free games, is *New World*, with a total of **one hundred million** (100000000) **players**. By consulting the auxiliary table, it was found that the *app_id* corresponds to an MMO (Massively Multiplayer Online) game, **where players cooperate but also compete against each other** (PVP), and **in-game transactions with real money are possible**.

The price of $39.99 falls within the expected range, as most games from well-known developers have a launch price of $69.99 for franchises already popular among players.

Player feedback for this game **is mostly positive** (154914) compared to negative (73900). After researching on Steam, it was found that the reason for **the negative feedback was due to in-game transactions disrupting the balance of player-versus-player fights** [1].

```
> jogo_mais_vendido_com_gratis
        name release_date price positive negative app_id max_owners
45483 Dota 2         2013     0  1477153   300437    570  200000000
```

The best-selling game, including free games, is *Dota 2*, with a total of **two hundred million** (200000000) **players**. By consulting the auxiliary table, it was found that the *app_id* corresponds to a game **where players cooperate** and where **in-game transactions with real money are possible**.

The game is free but includes in-game transactions and as **it is a game from 2013, it already has a well-established community that pays for exclusive in-game content**.

Player feedback for this game **is overwhelmingly positive** (1477153) compared to negative (300437). After researching on Steam, it was found that the reason for **the negative feedback was due to the game's steep learning curve and the presence of players with inappropriate behaviour in the community** [2].

4.2 – What is the highest-priced game on the platform?

Knowing which game has the highest price on the Steam platform is important to understand the price range of games available. This can indicate whether there is a demand for premium games and whether players are willing to pay more for certain titles. Additionally, this information can be useful for consumers who wish to make informed decisions about their game purchases.

```
> maior_preco
                        name release_date price positive negative app_id max_owners
32856 Aartform Curvy 3D 3.0         2013 299.9       32       13 253670      20000
```

The highest price on the Steam platform is $299.9, and it belongs to Artform Curvy 3D 3.0. Consulting the supporting table, the app_id corresponds to single-player games. As a result, I researched and found that **it is not a game but a sculpting program for 3D artists**. It is evident that this software has been around for over 10 years (**since 2013**), yet it has few users (max_owners=20000). **The feedback is mostly positive** but small in volume (32 positive, 13 negative). The reason for **the negative feedback is the high price in relation to the limited tools available in the program** [3].

4.3 – Which game has the most positive feedback?

Identifying the game with the most positive feedback can be helpful for players looking for high-quality gaming experiences. Additionally, for game developers, understanding what players value and appreciate in a game can guide the development of future projects and improve the overall quality of games released. This information can also be valuable for publishers and investors when assessing the potential success of a particular game.

```
> mais_feedback_positivo
       name release_date price positive negative app_id max_owners
45483 Dota 2         2013     0  1477153   300437    570  200000000
```

The game with the most positive feedback is Dota 2, with nearly one and a half million positive reviews. Dota 2 has already been mentioned in this project, as it also has the largest number of users (max_owners). This overwhelmingly positive feedback is due to it being a game with over 10 years of existence, during which it has built a strong community. Even after all this time, it continues to receive frequent updates. As a **free-to-play game**, it **allows many players to try it out**, enjoy the experience, and **contribute to the feedback** [2].

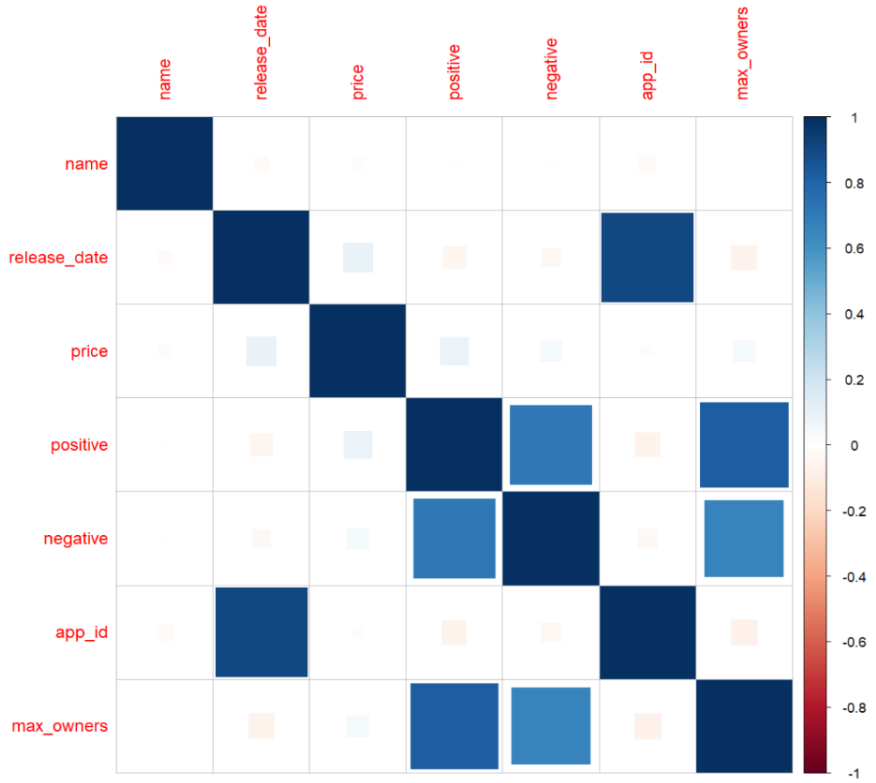
5 – Business Analytics Questions

5.1 – Data Correlation Analysis

Note: During the data cleaning process, two factors (minimum number of users and time to complete the game) were removed due to being incomplete and irrelevant for the analyses to be conducted.

To gain an overview of the data, I created a table and a correlation chart using the "square" method:

	name	release_date	price	positive	negative	app_id	max_owners
name	1.0000000000	-0.02096750	0.01472169	-0.005848503	-0.003741659	-0.02871208	0.0004475138
release_date	-0.0209675036	1.00000000	0.09592200	-0.055765131	-0.032913497	0.90415680	-0.0679578638
price	0.0147216862	0.09592200	1.00000000	0.081686196	0.046096420	0.01237417	0.0498768125
positive	-0.0058485029	-0.05576513	0.08168620	1.000000000	0.714050656	-0.06448060	0.8261974524
negative	-0.0037416593	-0.03291350	0.04609642	0.714050656	1.000000000	-0.03841487	0.6601004334
app_id	-0.0287120840	0.90415680	0.01237417	-0.064480599	-0.038414875	1.00000000	-0.0720966870
max_owners	0.0004475138	-0.06795786	0.04987681	0.826197452	0.660100433	-0.07209669	1.0000000000



By analysing the table and the data correlation chart:

1. **The release date (release_date) and game categories (app_id)** show the highest correlation of 0.904, within a range of 0 to 1.
2. **The maximum number of users (max_owners) and positive feedback (positive)** show a correlation of 0.826.
3. **Negative feedback (negative) and positive feedback (positive)** show a correlation of 0.714.
4. **The maximum number of users (max_owners) and negative feedback (negative)** show a correlation of 0.660.

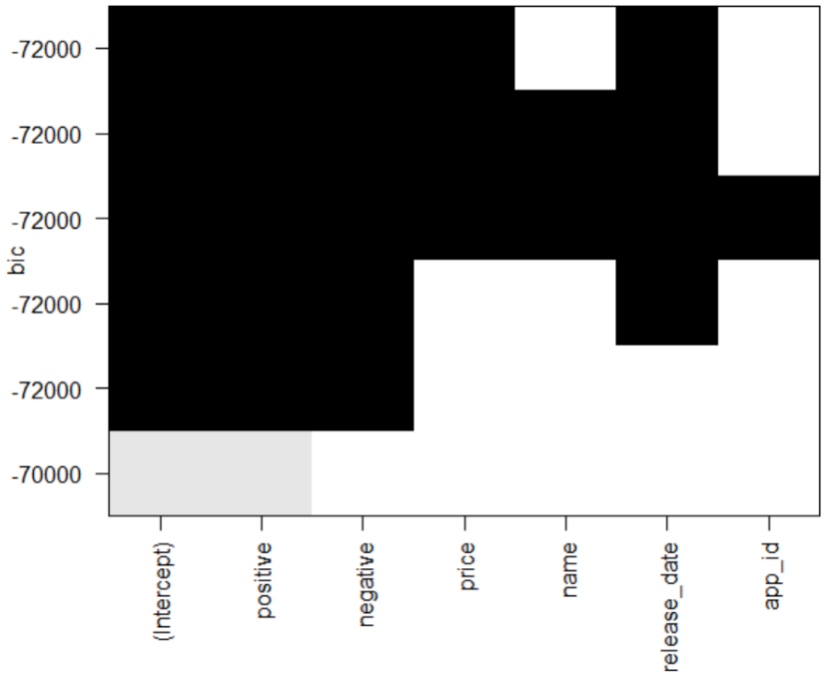
5.2 – What are the factors that most affect a game's sales?

This question is crucial for understanding the main drivers behind game sales on the Steam platform. By analysing various factors such as price, user reviews, release date, game genre, among others, it is possible to identify which of these factors have the greatest impact on sales. This can help developers, publishers, and investors make strategic decisions, such as setting appropriate prices, planning marketing campaigns, and guiding the development of new games.

Note: The dataset does not contain a *sales* column. To estimate sales, the maximum number of users (*max_owners*) was used, as these players had to acquire the games to play them.

For the analysis of the factors that most affect game sales, the fields with the highest correlation to the *max_owners* field were examined using a correlation table, and the best subsets were identified through a graphical representation.

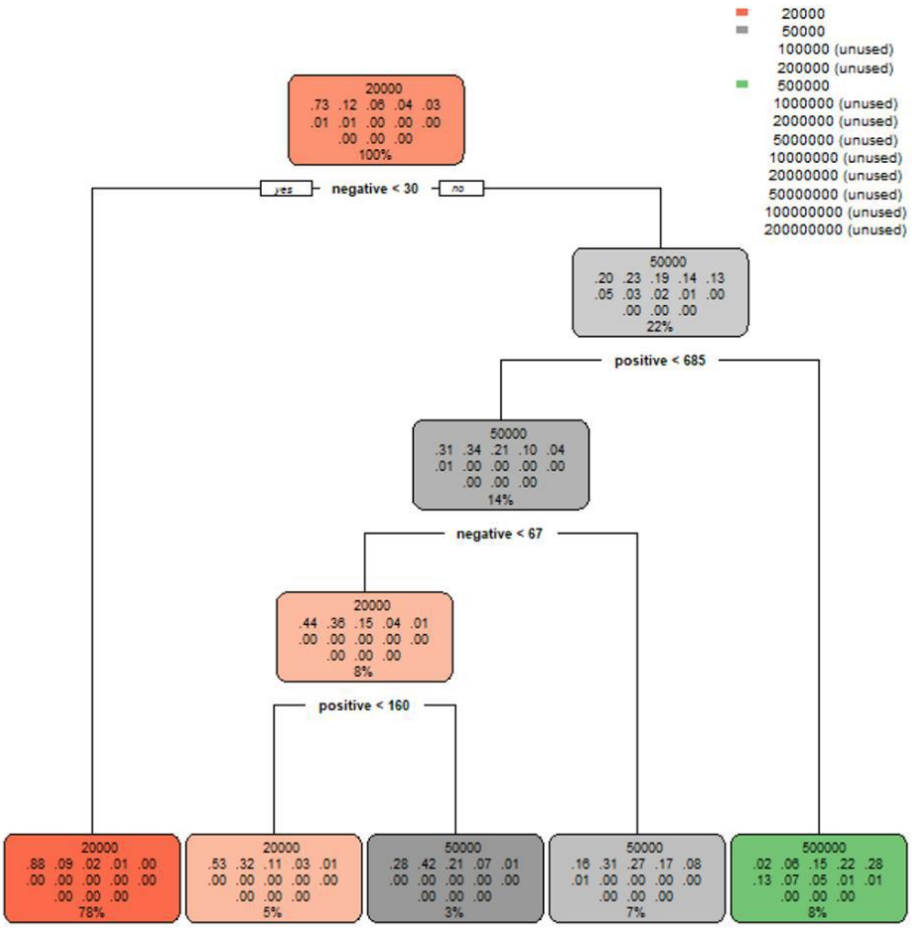
max_owners	positive	negative	price	name	release_date	app_id
1.0000000000	0.8261974524	0.6601004334	0.0498768125	0.0004475138	-0.0679578638	-0.0720966870



Using the table and the graph, it is observed that the fields with **the highest correlation to the max_owners field are positive and negative feedback**. Therefore, the factors that may influence this feedback will be analysed later. Additionally, **other fields of some importance include the release date (release_date) and the price (price)**.

The reasons for the high correlation between positive feedback and the number of users, followed by the correlation with negative feedback, are as follows:

- **Game Quality:** Games that receive positive player reception tend to attract more users. If a game is well-received by the gaming community due to its gameplay, graphics, storyline, or other elements, it is more likely to draw a larger user base.
- **Recommendations and Reviews:** Players often make purchasing decisions based on reviews and recommendations from other players. If a game receives positive feedback, it is more likely to attract new users who trust these reviews.
- **User Retention:** Games with positive feedback are more likely to retain existing users for longer periods. Player satisfaction keeps them engaged, leading to a larger user base over time.
- **Buzz Marketing (Word of Mouth):** Games with positive feedback are more likely to generate positive word-of-mouth marketing, where players recommend the game to friends and family. This can result in an increase in the number of users as more people become aware of and interested in the game.



While there is a correlation between positive feedback and the number of users, other factors not included in the analysis may also influence the user base. Additionally, the correlation between negative feedback and the number of users may indicate that games with low negative feedback also tend to attract more users (as observed in the decision tree, where 78% of users choose a game when it has low negative feedback).

5.3 – According to a set of factors, will the game receive positive feedback?

This question aims to understand the key factors that influence users' positive feedback on a game. By analysing a combination of factors such as price, user ratings, game genre, game duration, and others, it is possible to determine which characteristics lead to more positive feedback. This can be valuable for game developers, guiding the development process to create games that meet players' expectations and preferences, resulting in better reception and market success.

For the analysis of the factors that contribute to predicting whether a game will receive positive feedback, the fields with the highest correlation to the "positive" variable were examined using a correlation table.

positive	max_owners	negative	price	name	release_date	app_id
1.000000000	0.826197452	0.714050656	0.081686196	-0.005848503	-0.055765131	-0.064480599

This correlation table once again confirms the previously observed information, showing that positive feedback has a strong correlation with the **max_owners** field and negative feedback. The next most relevant factor is **price**.

Following this, a predictive model was built, which resulted in an R^2 value of:

```
Call:
lm(formula = positive ~ max_owners + negative + price, data = trainx)

Residuals:
    Min      1Q  Median      3Q     Max
-429060    -227      -3     119  468900

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.653e+02  4.572e+01  -5.802 6.61e-09 ***
max_owners   3.581e-03  4.070e-05  87.994  < 2e-16 ***
negative     3.041e+00  2.572e-02 118.240  < 2e-16 ***
price        3.495e+01  3.679e+00   9.500  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7115 on 39996 degrees of freedom
Multiple R-squared:  0.754,     Adjusted R-squared:  0.754
F-statistic: 4.087e+04 on 3 and 39996 DF,  p-value: < 2.2e-16
```

A **coefficient of determination (R^2) of 0.754** indicates that approximately **75.4% of the variability** in the response variable (**positive feedback**) can be explained by the model. This means that the model can capture and explaining a significant portion of the observed variability in players' positive feedback regarding games.

An R^2 of 0.754 is considered quite high and suggests that the model has a reasonably strong ability to predict positive feedback based on the given data. This can be interpreted as an indication that the factors considered in the model (such as negative feedback, price, etc.) have a strong relationship with players' positive feedback.

However, it is important to emphasize that **R^2 alone does not determine the validity or quality of the model**. Therefore, it is always crucial to conduct a more in-depth analysis, including the evaluation of other model performance metrics and the consideration of potential limitations in the data and modelling.

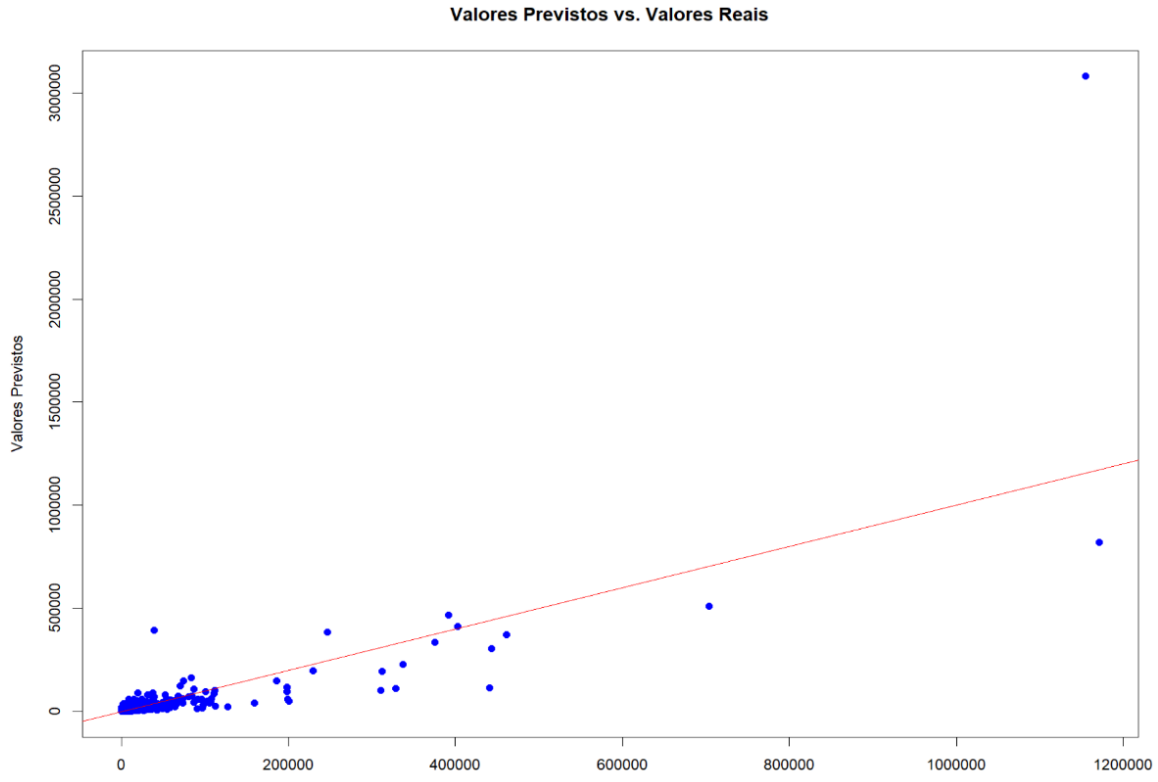
To support this evaluation, I created a **table and a scatter plot** comparing the predicted values to the actual values:

- **Table with actual values:**

```
> summary(steamgames)
      name         release_date         price            positive          negative            app_id          max_owners
 Min.   :    1   Min.   :2013   Min.   :  0.000   Min.   :      0   Min.   :     0.0   Min.   :    570   Min.   :    20000
 1st Qu.:14484   1st Qu.:2018   1st Qu.:  1.990   1st Qu.:      4   1st Qu.:     1.0   1st Qu.: 676023   1st Qu.:    20000
 Median :29593   Median :2020   Median :  4.990   Median :     16   Median :     4.0   Median :1095830   Median :    20000
 Mean   :29602   Mean   :2019   Mean   :  7.815   Mean   :   1046   Mean   :   193.5   Mean   :1165656   Mean   :   135329
 3rd Qu.:44690   3rd Qu.:2021   3rd Qu.:  9.990   3rd Qu.:     80   3rd Qu.:    24.0   3rd Qu.:1579013   3rd Qu.:    50000
 Max.   :59798   Max.   :2023   Max.   :299.900   Max.   :1477153   Max.   :895978.0   Max.   :2690780   Max.   :200000000
```

- **Table with predicted values:**

```
     price              positive            negative           max_owners          predictions
 Min.   :  0.000   Min.   :      0   Min.   :     0.0   Min.   :    20000   Min.   :  -193.6
 1st Qu.:  1.990   1st Qu.:      4   1st Qu.:     1.0   1st Qu.:    20000   1st Qu.:   -75.4
 Median :  4.990   Median :     16   Median :     4.0   Median :    20000   Median :    82.7
 Mean   :  7.853   Mean   :   1058   Mean   :   225.6   Mean   :   129946   Mean   :  1160.5
 3rd Qu.:  9.990   3rd Qu.:     81   3rd Qu.:    24.0   3rd Qu.:    50000   3rd Qu.:   366.8
 Max.   :299.900   Max.   :1171197   Max.   :895978.0   Max.   :100000000   Max.   :3082297.0
```



By comparing the predicted **positive feedback** values with the actual values in the original dataset, it is evident that the values are **close**. The **minimum** value for both the actual and predicted feedback is **0**. The **maximum actual** positive feedback value is **1477153**, while the **predicted maximum** is **1171197**, resulting in a difference of **305956** between the actual and predicted maximum values. The **average actual** value is **1046**, whereas the **predicted average** is **1058**.

By analysing the scatter plot, it is evident that the predicted values and actual values are close. However, some values are more dispersed and not as condensed.

Therefore, this model is good for predicting positive feedback for games, considering only that the maximum predicted value may not be as close to the actual value, tending to be lower than the real value.

It is concluded that **a game will receive positive feedback if**:

- **It has many players**, as the game will receive more positive feedback. This happens if the game becomes popular, and users are satisfied with the overall gaming experience.
- **It receives less negative feedback**, suggesting an inverse relationship between user satisfaction and negative reviews.
- **Its price meets expectations**. Higher-priced games may be perceived as offering greater value and, therefore, receive more positive feedback, provided they meet user expectations.

6 – Share

The R script used in this project can be downloaded here:

- Click here to view the R file

This file contains all the code used for data preparation, analysis, and visualization, allowing for review or adaptation as needed for future analyses.

7 – Conclusion

With this R programming project on Steam platform games from 2013 to 2023, it was possible to gain valuable insights into the behaviour of both games and users. The selected questions for analysis were fundamental in understanding various aspects of the online gaming market.

We began by identifying the best-selling game over the years, considering both paid and free games. Then, we determined the highest-priced games on the platform and identified the game with the most positive feedback, which allowed us to better understand player preferences.

One of the most important analyses was identifying the factors that most affect a game's sales. By exploring the correlation between different variables, such as positive feedback, negative feedback, price, and others, we were able to identify significant patterns that influence a game's commercial success on the platform.
An additional question raised was whether, based on a set of factors, it would be possible to predict whether a game would receive positive feedback.

Despite the challenges encountered, such as the need for prior dataset cleaning (over 60,000 records cleaned and normalized in Excel, Power BI, and R) and the selection of appropriate models, including the use of *rpart* for decision tree construction instead of *C50* and the analysis of R^2 and scatter plots (without using the ROC model) to assess the quality of the prediction model, we managed to reach relevant conclusions.

In conclusion, this project provided a comprehensive view of the Steam gaming market, demonstrating the importance of data analysis in understanding and decision-making in the digital entertainment sector. The findings and methodologies employed in this project can serve as a foundation for future analyses and for improving game development and commercialization strategies on the platform.

8 – Bibliography

[1] Valve Corporation, *New World*, consulted on Abril 6, 2024:
https://store.steampowered.com/app/1063730/New_World/

[2] Valve Corporation, *Dota 2*, consulted on Abril 6, 2024:
https://store.steampowered.com/app/570/Dota_2/

[3] Valve Corporation, *AartfromCurvy 3D 3.0*, consulted on Abril 6, 2024:
https://store.steampowered.com/app/253670/Aartform_Curvy_3D_30/