Universitat de les Illes Balears

21746 - Data Mining

Final Project

Steam Successful Indie Games Study

Iván Pulgar Rodas
Jordi Sevilla Marí
Nahuel Vázquez
Yelyzaveta Denysova
Xiaozhe Cheng
Gabriel Oliver Artigues
Arturo Mus Mejías

Contents

1 Introduction

Steam (https://store.steampowered.com/) is the largest digital distribution platform for PC games, hosting thousands of games at all prices. In this project, we will work on a large scraped dataset of Steam games (up to 2025) and try to understand two linked phenomena: (i) which patterns are common among successful indie games, and (ii) how game genres evolve over time (Lifecycle).

1.1 Explanation of the Dataset

The core data source is the *Steam Games Dataset* from Kaggle https://www.kaggle.com/datasets/artermiloff/steam-games-dataset/data, containing scraped information about games published on Steam up to 2025 (metadata, tags/genres, estimated owners, ...). We complemented and validated certain games using SteamDB (https://steamdb.info/), which provides additional public information and estimations.

The dataset contains a total of **94,948** observations and **47** variables. The data contains a mix of numeric values (price, reviews, owners estimates, playtime, ...) and textual/categorical descriptors (genres, tags, categories, developers, ...).

- **appid**: Unique identifier of the game on Steam. [num]
- **name**: Name of the game. [text]
- **release_date**: Represents the date where the game was released. [time]
- **required_age**: Minimum age required to play the game. [num]
- **price**: How much the game costs. If its 0 it means that the game is Free-to-Play. [num]
- **dlc_count**: Amount of DLCs (Downloadable Contents) the game has. [num]
- **support_url**: URL to the support page of the game. [text]
- **windows, mac, linux**: Platforms where the game can be run. [categorical]
- **metacritic_score**: Metacritic score based on professional reviews. [num]
- **achievements**: Number of achievements the game has. [num]
- **recommendations**: Number of user recommendations. [num]
- **supported_languages**: Languages supported by the game. [text]
- **packages**: Available packages for the game. It contains the name and a description of the package and the names, descriptions and subprices of the subpackages. [text]
- **developers**: Developers associated with the game. [text]
- **publishers**: Publishers associated with the game. [text]
- **categories**: Categories that the game belongs to. [text]
- **genres**: Genres that the game belongs to. [text]
- **positive**: Number of positive user reviews. [num]
- **negative**: Number of negative user reviews. [num]
- **estimated_owners**: Estimated number of owners. [num]
- **average_playtime_forever**: Average playtime since March 2009 measured in minutes. [num]
- **average_playtime_2weeks**: Average playtime in the last two weeks measured in minutes. [num]
- **median_playtime_forever**: Median playtime since March 2009 measured in minutes. [num]
- **median_playtime_2weeks**: Median playtime in the last two weeks measured in minutes.[num]
- **peak_ccu**: Number of current users playing the day before the data was scrapped. [num]
- **tags**: List of tags the game has with its name and its key. [text]
- **pct_pos_total**: Percentage of all reviews that are positive. [num]
- **num_reviews_total**: Number of the total reviews the game has. [num]

1.2 Objectives

The objective of this project is to analyse the Steam video game ecosystem with a focus on indie titles, using data mining and exploratory data analysis techniques to extract interpretable and actionable insights.

1. **Prepare and structure the Steam dataset for analysis** by identifying relevant variables, handling missing, duplicated, and inconsistent values, and formatting raw attributes ready for data mining techniques.
2. **Analyze long-term genre dynamics on Steam** by studying how grouped game genres evolve over time in terms of market share and player engagement.
3. **Identify common patterns among successful indie games** by constructing an operational proxy for success using clustering over engagement and market signals, and then mining frequent genre, mechanic, and characteristic combinations within the successful subset.
4. **Evaluate the role of pricing and visibility** by studying the relationship between price, estimated owners, and engagement indicators using multivariate regression.
5. **Study the influence of successful indie games on genre and tag evolution** through targeted case studies that analyze changes in tag usage before and after the release of highly influential titles.

The goal is not to establish rules for success, but to highlight recurring patterns and associations that can help better understand the indie game market and help game design decisions.

2 Processing the Data

Data preprocessing prepares raw data for analysis by addressing quality issues and transforming data into suitable formats for mining algorithms. After exploring the dataset and its attributes we performed an initial inspection that revealed data quality problems, including inconsistent formatting, missing values, and negative values among the whole dataset. To ensure the dataset was prepared for subsequent analysis we applied some preprocessing techniques described below.

2.1 Handling Irrelevant Attributes

As it was shown before, the dataset contains 47 variables, however, only the variables relevant to our analysis were retained and explained. The remaining variables were removed in order to reduce the dimensionality of the dataset (those being: *detailed_description*, *about_the_game*, *short_description*, *reviews*, *header_image*, *website*, *support_url*, *support_email*, *metacritic_url*, *notes*, *full_audio_languages*, *screenshots*, *movies*, *user_score*, *score_rank*, *discount*, *pct_pos_recent*, *num_reviews_recent*). It's also noticeable that there where one of the removed attributes containing missing values on all of its observations (*score_rank*). This could have been because some scraping error, but we removed that variable as it is not useful for our investigations.

2.2 Reducing Noise in the Dataset

Noise refers to data that does not provide meaningful or reliable information for analysis and may negatively affect the performance of the data mining algorithms. The Steam dataset contained several sources of noise, so we performed an investigation to identify these issues and determine how to handle them.

First, we identified that there were some games within the dataset that had their playtest version on it. These playtest observations do not contain relevant information, as they typically have no price, reviews, or meaningful data. Therefore those observations with the *Playtest* keyword were removed.

Additionally, we also found games with blank names, so they were also removed. Another interesting issue we found was the presence of multiple observations corresponding to the same game. We investigated whether these duplicated observations followed a pattern (such as representing the discounted versions or re-published editions), but we could not find a global correlation. In some cases, duplicates reflected different discount states, while in others they differed only in the total number of reviews. To solve this issue, we kept the version of each duplicated game with the highest price and the highest total number of reviews

2.3 Data Formatting and Addition

Initially, the *released_date* attribute was stored in text format. Therefore, we converted to R's `Date` format, and we removed those observations with a missing or invalid released date.

The *estimated_owners* attribute represents a range of values (e.g., 100-20.000). To make this information more suitable for analysis, we used a self made function to split this range into two new attributes: *estimated_owners_min* and *estimated_owners_max*.

2.4 Better Genre Classification

The original dataset have `genre` and `tag`, which tag also contains some genres and they are mixed up. So we have created our own classification of 14 genres and each genre have their own subgenres.

Genre	**Subgenres**
Action	Action, Action-Adventure, Action RPG, ...
Shooter	Arena Shooter, Battle Royale, ...
Roguelike	Rogue-like, Rogue-lite, Roguevania, ...
RPG	RPG, JRPG, CRPG, ...
Strategy	Strategy, Grand Strategy, RTS, ...
Simulation	Simulation, Automobile Sim, Farming Sim, ...
Sport	Sports, Baseball, Basketball, ...
Fighting	2D Fighter, 3D Fighter, Souls-like, ...
Platformer	Platformer, 2D Platformer, 3D Platformer, ...
Tabletop	Puzzle, Puzzle-Platformer, Logic, ...
Casual	Casual, Idler, Clicker, ...
Horror	Horror, Psychological Horror
Music	Music, Rhythm, Typing
Walking_sim	Walking Simulator, FMV

Table 1: Game Genres and Associated Subgenres

The complete structure of the genres is in the original `Rmd`.

3 Exploratory Data Analysis

How people play games have been changing through time, so we found that it would be interesting to see how the video game have been evolving and see how are the players of each genre.

In this section, we will be focusing on studying the behavior of the video games genre, specifically the market share evolution and the playtime of the players.

3.1 Genre Market Share Trends

In order to obtain an objective view of whether a genre has truly grown over time, we will measure the market share of each genre of each year. The main advantage of using this approach is that the grown of each year is relative to the total grown of the game industry. Using the raw count of released games would not allow us to distinguish whether the game popularity has actually grown or if it is because the general game market has a growing tendency. The equation we will be using is at it follows:

$$\text{Market Share} = \frac{\text{Genre Games}}{\text{Total Games}}$$

We now define some metrics to evaluate the performance and behavior of the genres. *Net Growth* is defined as the difference between the final and the initial market share, and, with this metric, we can identify which genre has grown the most. Another metric is *Volatility*, which is computed as the sum of all the absolute differences between the previous and current year.

$$\text{Net Growth} = \text{Market Share}_{\text{final}} - \text{Market Share}_{initial} \tag{1}$$

$$\text{Volatility} = \sum |\Delta\text{Market Share}| \tag{2}$$

Once we computed this formulas, we will first plot the genres by *Net Growth*.

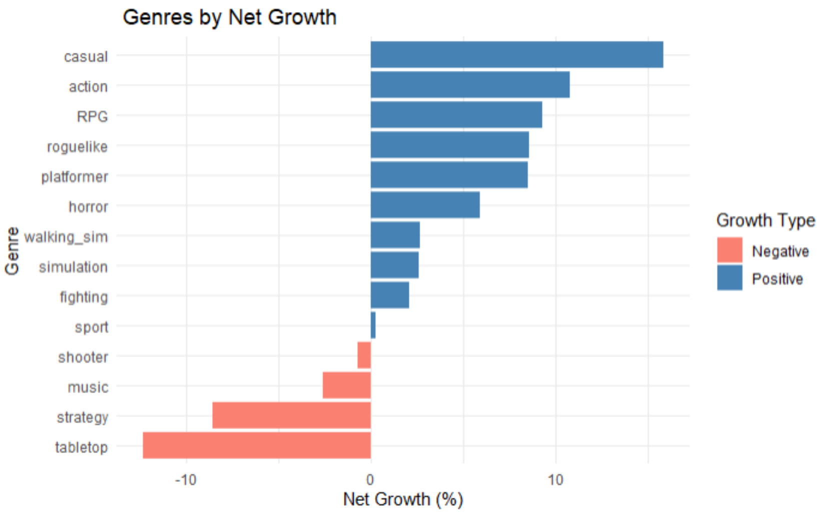


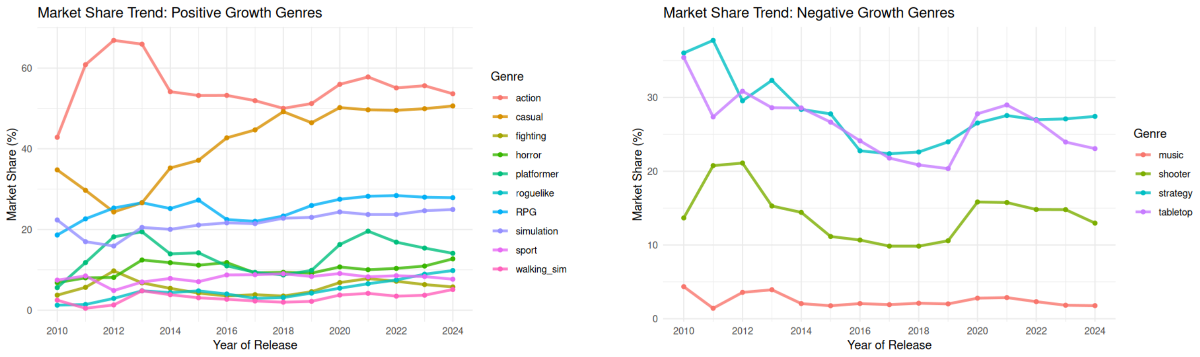
Figure 1: Top and bottom net growth

We observe a rise of the casual genre. This can be explained by the fact that video games have become more popular and are no longer a niche hobby. It has also evolved into a social activity, particularly among young people, who often use chat platforms such as Discord to play together. This helps explain the growth of casual multiplayer games such as "Fall Guys", "Lethal Company", "R.E.P.O.", "PEAK" and "MIMESIS".

Genres such as action, RPG, roguelike and platformer also appear at the top. These genres are commonly part of indie games, and since indie games occupies the majority of the game publication, it is expected that they they are on the top. An interest thing is the growth of the horror. Although horror is often considered a niche category, we think its rise may be partially driven by the stream culture, as horror games tend to engage more when played by streamers, increasing their visibility and popularity.

On the other hand, the genres belonging to the bottom 3 are music, strategy and tabletop games, which are more niche and appeal to smaller audiences. A particularly interesting case is the shooter genre, which has a negative grown. This is interesting because they player-base is often very wide and the shooter genre is one of the most viewed category of Twitch. One possible explanation could be that the shooter genre is highly concentrated around a small number of long-lasting and highly polished competitive titles (e.g., CS:GO/CS2, PUBG, Rainbow Six Siege, Apex Legends, Overwatch). Due to the competitive nature of these games, players tends to focus on the same games in order to have a metric and show their skills. So despite its large community, the shooter genre does not grow in term of Market Share, as relatively few new titles gain traction.

Now let's see the evolution of the market share through time.



(a) Positive Trend

(b) Negative Trend

Figure 2: Comparison of Market Share Trends

The top 5 most volatile genres using the previous volatile definition.

Genre	**Volatility**
action	53.97532
casual	43.56214
platformer	41.42804
tabletop	36.58252
strategy	28.80528

Table 2: Genre Volatility

We can see reflected in the graph the most volatile genre like action which in 2012 peaked, casual which is in constant rising and platformer which peaked in 2013 then it went down and later in 2021 it peaked again.

Tabletop games shows an overall decreasing trend, with a recovery between 2020 and 2022. Finally, the strategy genre seems to be recovering since 2016 and it shows signs of stabilization towards 2024.

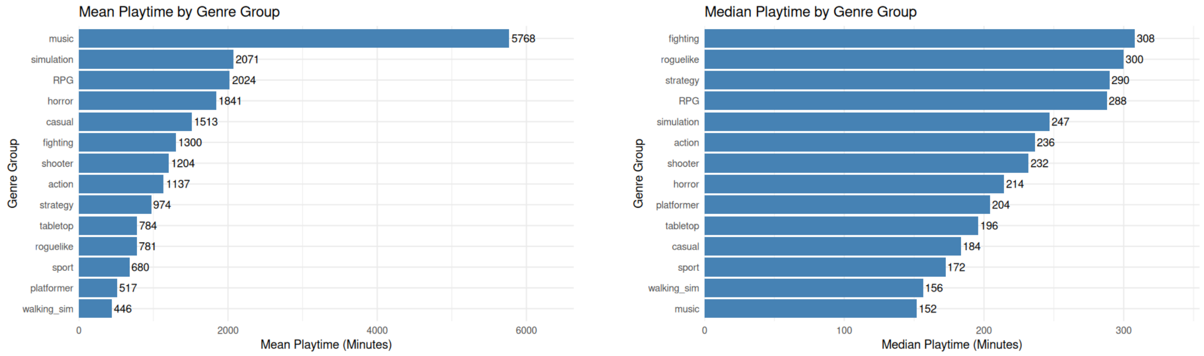
3.1.1 Conclusion

In conclusion, the video game genre market is difficult to predict, as it is mainly driven by player preferences and trending. Genres that are initially niche can become very popular due to influences, such as a vety popular content creators playing them or the release of a highly successful title that reshapes its genre, such as "Holow Knigh" to metroidvania genre and "The Binding of Isaac" to the roguelike genre (studied in section 5).

Another clear example of external influence is the COVID-19 pandemic. As in-person social activities were restricted, many players turned to virtual board or card games, such as "UNO" or the most popular "Tabletop Simulator", a collection of all types of Board, Card, Roleplay games, that also gave the ability to create their own ones.

3.2 Genre Player Playtime

We mow analyze how player playtime behaves across different genre group. We will be using the mean playtime to have an idea of which genres accumulate the highest total playtime. In addition,the median playtime is used to have a view of a typical player playtime.



(a) Mean

(b) Median

Figure 3: Mean and Median

Figure 3 shows that the mean playtime is a bigger than the median playtime, indicating that the playtime distribution of playtime is skewed. This skewness is mainly because a small group of "hardcore" players that spends a much higher quantity of time than the majority of user, which make the mean higher.

To investigate which genres contains the more hardcore players, we compute a skewness ratio for each genre by dividing the mean playtime by the median playtime. If the skewness ratio is bigger than 1 it indicates that the genre contains "hardcore" players that affects the mean by increasing its value.

$$\text{Skewness Ratio} = \frac{\text{Mean}}{\text{Median}}$$

Visualization of the skewness ratio.

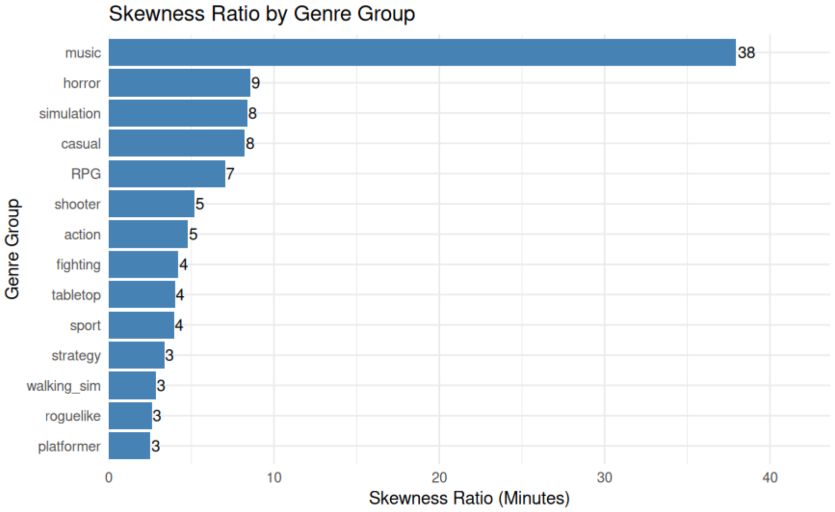


Figure 4: Skewness Ratio

We can see that all genres have $skewness\_ratio \geq 3$, indicating that, by nature, tall the games will have a small group of "hardcore" players that will have a lot more playtime than the majority of the players. It is interesting to point that music genre stand out with a much higher skewness ratio compared to the others. In fact, music have the top 1 mean playtime and the top 10 median playtime. This suggests that this genre is very niche, while most of the players will just give it a try and wont spend that much time playing the "hardcore" players will spend a huge amount of time.

4 Commonalities among successful indie games

Indie game development is high-risk: budgets are limited, marketing reach is uncertain, and audience discovery can be unpredictable. This section explores commonalities across successful indie games, with two practical goals:

- **Understand market audience**: Identify what combinations of genres/mechanics tend to co-occur among games that reach larger audiences.
- **Reduce risk for game-making**: Extract patterns that can guide design decisions without prescribing a single “correct” formula.

In this section we will be focusing on four questions:

- **Does genre matter?**
- **Do mechanics matter (especially in combination with genre)?**
- **Do game characteristics matter (e.g., camera, player modes, VR)?**
- **Does pricing matter (relationship with owners)?**

4.1 Selecting indie games

Steam has done the hard work for us by including *Indie* as a genre/tag (and related tags such as *Crowdfunded / Kickstarter*). Since these tags have been assigned by users worldwide, we can agree on these tags representing widely-considered indie games. Therefore, we filter the dataset to retain games containing any of these indicators in their Genres, Tags or Categories.

4.2 Determining *successful* indie games

Since *success* is not explicitly labeled in the dataset, we have build an operational proxy based on market and engagement signals. The main goal of this step is to isolate a subset of indie games that consistently outperform the rest, ensuring that later analyses (such as association rules and plots) focus on patterns that appear among higher-impact titles.

4.2.1 Clustering approach

We cluster indie games using a set of numeric variables that capture outcomes and engagement (e.g., recommendations, reviews, playtime, peak CCU, owners/revenue estimates, price, and platform availability). Before fixing the final configuration, we experimented with multiple numbers of clusters: **2 clusters** as we were attempting to represent successful vs not successful; **3 clusters** as we were attempting to represent massively successful, successful and not successful; **4 clusters**; **5 clusters**; **10 clusters**

The most interpretable segmentation that we achieved was using **3 clusters**.

In our dataset, representative examples of these mid/high clusters included games such as Deceit, Graveyard Keeper, Unturned (mid-tier), and Stardew Valley, Subnautica, Terraria (top-tier). These examples illustrate how the clustering captures meaningful outcome tiers rather than arbitrary partitions.

Based on this results, we implemented the final clustering using the k-means algorithm with $k = 3$ on standardized variables. Using this operational definition, we obtained a total number of 37,500 games successfully labeled (57.8% of the indie games).

4.2.2 Validation: do clusters separate meaningfully?

Since clustering is an unsupervised technique, we validate the quality using a simple but practical hypothesis: **if the most discriminant variable is unable to separate the clusters, then the clustering is unlikely to be meaningful**. Therefore compute the variable that best differentiates the clusters by computing the highest ANOVA F-statistic and examine wether it produces a clear separation across the groups.

In our case, the most discriminant variable is *pct_pos_total* and it produces a clean ordering between the three clusters, which supports using this partition in the rest of the report.

As an additional sanity check, we compute a global silhouette score on a random sample of games in the standardized feature space. Silhouette values closer to 1 indicate well-separated clusters, values near 0 indicate overlapping clusters, and negative values suggest poor assignments. The resulting average silhouette score on the sample is 0.395, indicating a moderate yet meaningful cluster separation.

4.2.3 Cluster interpretation

The figure below summarizes the mean and median values of the discriminant variable (*pct_pos_total*) across clusters, alongside the number of games per tier.

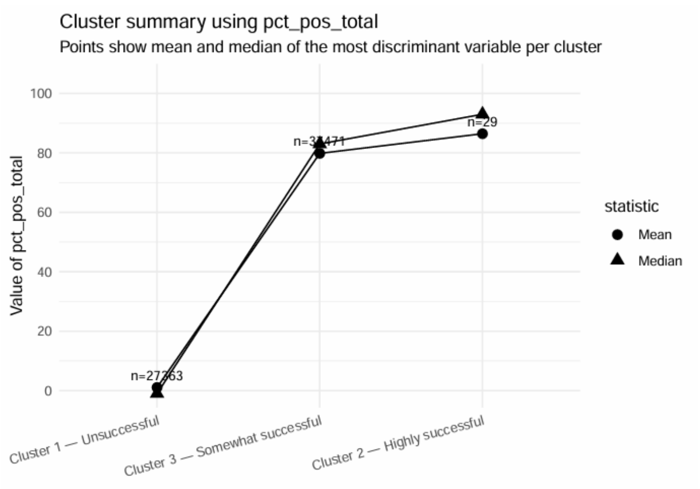


Figure 5: Cluster summary using the discriminant variable

In our run, the discriminant variable is *pct_pos_total*, which acts as a strong quality/visibility signal: the unsuccessful long-tail cluster concentrates near the lowest values, while the two top-tier cluster achieves the highest values, separating games with limited traction from games with stable success. A key takeaway from the cluster sizes is that the distribution is highly imbalanced: a large share of games are concentrated in the lower-to-mid tiers, while the top tier contains only a small number of games. In this run, the lowest tier accounts for , whereas the highest tier represents only of indie games.

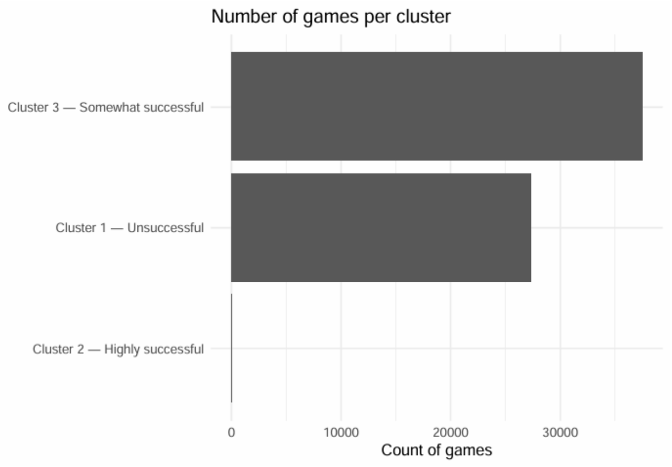


Figure 6: Number of games in each cluster

Since the clusters are interpretable and separable (both visually and by the discriminant variable test), we use them to define a successful subset and proceed with pattern mining in the next sections.

4.3 Genre study

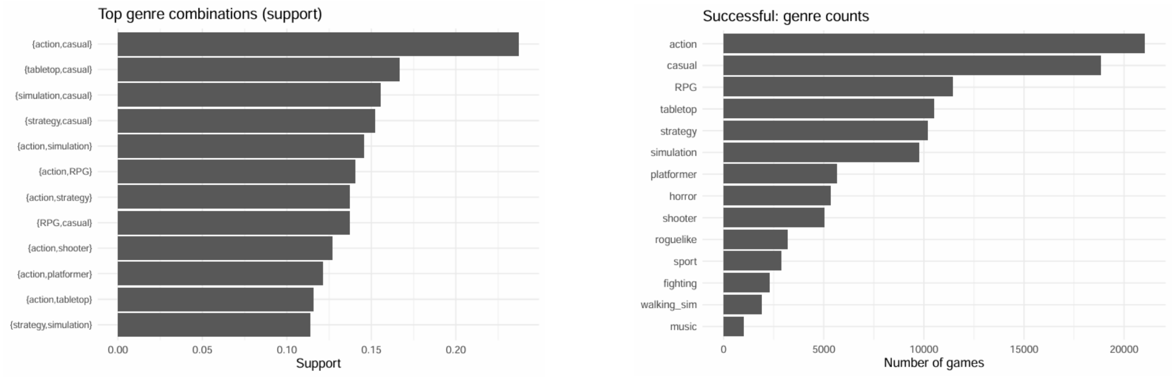
As we explained in section 2.4 Steam genre labels are sometimes inconsistent, therefore we use a grouped genre taxonomy specifically created for this project (e.g., action, RPG, strategy, platformer, etc.). Each successful game is mapped into one or more of these grouped genres using its Genres, Tags, Categories, as described previously.

Before adopting this approach, we initially attempted to mine association rules that directly predicted success by fixing the right-hand side of the rules to *successful = TRUE*. However, this strategy produced obscure and highly specific combinations with limited generality: many rules were driven by niche tag mixes or small subsets of games, and they did not reflect the kinds of genre patterns seen in widely recognized successful indie titles.

For that reason, we switched to a more interpretable strategy. Rather than predicting success directly, we first filter the dataset to successful indie games, and then mine frequent item sets (genre combinations) with sufficient support inside this subset. This approach still answers our question: **Among successful games, what genre combinations are most common?**

4.3.1 Most common genre combinations (frequent itemsets)

The table below lists the most frequent genre combinations and genre prevalence among successful games (support = proportion of successful games containing that combination).



(a) Most frequent Genre Combinations

(b) Genre Prevalence

Figure 7: Most Frequent Genre combinations and Genre Prevalence among successful indie games

4.3.2 Conclusions from genres

The grouped genre counts show a highly concentrated distribution: successful indie games most frequently fall under action and casual, followed by a second tier including RPG, tabletop/puzzle, strategy, and simulation. This suggests that, within the successful segment, many games belong to genres that are either broad and audience-friendly (action/casual), or built around deep progression and replayability (RPG/ strategy/ simulation/ tabletop).

Looking at frequent genre combinations (Eclat), the strongest result is the pairing action, casual, which appears in roughly one quarter of successful games. More generally, casual acts as a “bridge” genre: it appears in many of the top combinations (e.g., tabletop + casual, simulation + casual, strategy + casual, and RPG + casual). This indicates that many successful indie games mix a core genre identity with accessible design traits (short sessions or low entry difficulty).

The second major pattern is that action combines well with several popular genres: action + RPG, action + simulation, action + strategy, action + shooter, and action + platformer all appear as frequent itemsets. This highlights a typical indie design strategy: start from a strong core action loop and enrich it with complementary systems such as progression (RPG), management (simulation/strategy), or movement challenges (platformer).

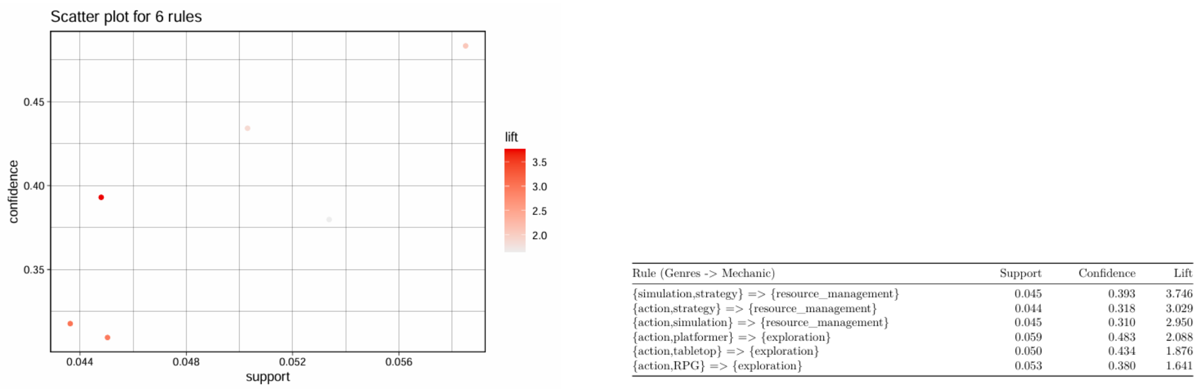
4.4 Mechanics study

Mechanics are stored as a list of tags (e.g., resource_management, procedural, narrative, card/deckbuilding, etc.) in the steam dataset so they are first one-hot encoded in a binary manner so the Apriori Algorithm can be executed over them. Instead of only asking "which mechanics are common?", we also ask: **Given a genre combination, which mechanics are most strongly associated with it?**

In order to answer the question we will be mining association rules of the form:

$$LHS(antecedent) : genregroup(s) - > RHS(consequent) : mechanicgroup$$

Rules are ranked by *lift* to highlight mechanics that occur more often than expected within a genre context. The *support* is set to *0.01* and *confidence* to *0.3*. Higher values fail to find interesting relationships in the data.



(a) Scatter plot of the rules

Rule (Genres -> Mechanic)	Support	Confidence	Lift
{simulation,strategy} => {resource_management}	0.045	0.393	3.746
{action,strategy} => {resource_management}	0.044	0.318	3.029
{action,simulation} => {resource_management}	0.045	0.310	2.950
{action,platformer} => {exploration}	0.059	0.483	2.088
{action,tabletop} => {exploration}	0.050	0.434	1.876
{action,RPG} => {exploration}	0.053	0.380	1.641

(b) Table of the association rules obtained

Figure 8: Results obtained by applying the association rules technique

4.4.1 Conclusions from genres and mechanic rules

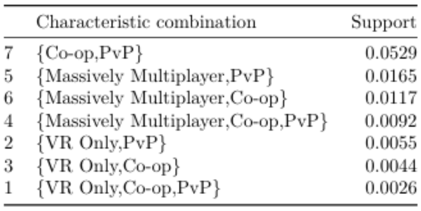
A first clear pattern is the strength of resource management inside strategy/simulation hybrids. The rule $\{simulation, strategy\} - > \{resource\_management\}$ has the highest lift (around 3.75), meaning that management-oriented mechanics are several times more likely than expected when a game sits at the intersection of these genres. Similar high-lift rules also appear for $\{action, strategy\}$ and $\{action, simulation\}$, suggesting that successful hybrids often combine an action layer with systems such as crafting, automation, building, or economy loops.

A second pattern is that exploration frequently complements action-driven genres. Rules such as $\{action, platformer\} - > \{exploration\}$, $\{action, tabletop\} - > \{exploration\}$, and $\{action, RPG\} - > \{exploration\}$ show relatively high confidence, indicating that when successful games blend action with movement/progression, they often reinforce the experience through discovery loops (new areas, loot, dungeons, collectables, or open-ended progression paths).

Overall, the results support a practical conclusion: successful indie games often rely on a core genre identity and then amplify engagement through a matching mechanic; management systems for strategy/simulation hybrids, and exploration loops for action-driven combinations. These associations do not prove causality, but they highlight combinations that repeatedly co-occur among successful titles and can reduce design risk by aligning with common audience expectations.

4.5 Game characteristics study

Characteristics are captured from Steam categories/tags that describe presentation and play modes (e.g., 2D/3D, first-person/third-person, single-player/co-op, VR). We mine frequent characteristic combinations using Eclat.



	Characteristic combination	Support
7	{Co-op,PvP}	0.0529
5	{Massively Multiplayer,PvP}	0.0165
6	{Massively Multiplayer,Co-op}	0.0117
4	{Massively Multiplayer,Co-op,PvP}	0.0092
2	{VR Only,PvP}	0.0055
3	{VR Only,Co-op}	0.0044
1	{VR Only,Co-op,PvP}	0.0026

Figure 9: Characteristic combinations with highest support

4.5.1 Conclusions from characteristics

The most frequent characteristic combinations in the successful subset are primarily multiplayer-focused, especially the pairing $\{Co - op, PvP\}$ ($support \approx 0.053$). This suggests that, among successful indie games that include multiplayer features, a common design choice is to combine collaboration and competition within the same title.

We also observe VR only appearing in the top combinations but with low support, suggesting that VR-exclusive successful games exist but represent a niche segment compared with traditional PC titles. A plausible explanation is that the VR market offers fewer alternatives overall, so the relatively small number of VR-only titles can capture a larger share of VR players and reach the engagement thresholds needed to be labeled as successful in our proxy, which makes VR-only features show up among the top itemsets.

Overall, characteristics appear to matter mainly as experience modifiers: multiplayer modes (co-op/PvP) are recurring patterns among successful games, while more specialized formats (like VR-only) are less common.

4.6 Pricing and Owners

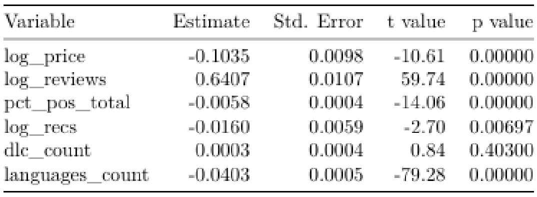
Finally, we explore whether price is related to estimated owners. We initially attempted to approach this with a simple scatter plot which ended up hard to interpret due to heavy overlap and the long-tail nature of owners. Instead the best analysis technique has been running a multivariate regression on non-free games and studying the weights.

4.6.1 Multivariate regression

A raw relationship between price and owners can be misleading because both variables are correlated with other factors. To reduce this bias, we estimate a multivariate regression on paid games only (price > 0), controlling for: **Review volume** (additional approximate for audience size); **Rating quality** (proxy for perceived value); **Recommendations** (proxy for engagement); **DLC count and language count** (rough indicators of production scope).

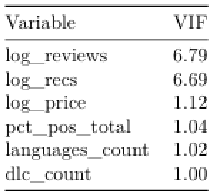
The coefficient of log-price can be interpreted as the partial association between price and owners after accounting for these variables.

In order to ensure the relevance of the analysis to the indie game market, the price has been capped at 90 euros and games above that price have been removed as they do not represent realistic pricing representations of indie games.



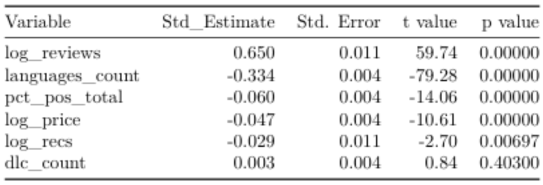
Variable	Estimate	Std. Error	t value	p value
log_price	-0.1035	0.0098	-10.61	0.00000
log_reviews	0.6407	0.0107	59.74	0.00000
pct_pos_total	-0.0058	0.0004	-14.06	0.00000
log_recs	-0.0160	0.0059	-2.70	0.00697
dlc_count	0.0003	0.0004	0.84	0.40300
languages_count	-0.0403	0.0005	-79.28	0.00000

(a) Multivariate regression coefficients(paid games only)



Variable	VIF
log_reviews	6.79
log_recs	6.69
log_price	1.12
pct_pos_total	1.04
languages_count	1.02
dlc_count	1.00

(b) Variance Inflation Factors(VIF) for multicollinearity diagnosis



Variable	Std_Estimate	Std. Error	t value	p value
log_reviews	0.650	0.011	59.74	0.00000
languages_count	-0.334	0.004	-79.28	0.00000
pct_pos_total	-0.060	0.004	-14.06	0.00000
log_price	-0.047	0.004	-10.61	0.00000
log_recs	-0.029	0.011	-2.70	0.00697
dlc_count	0.003	0.004	0.84	0.40300

(c) Standardized regression coefficients

Figure 10: Summary of multivariate regression diagnostics, including raw coefficients, VIF values, and standardized effects for paid indie games.

The *Variance Inflation Factors* (VIF) table provides an explicit check for multicollinearity. When *VIF* values are elevated, predictors overlap strongly (for example, review volume and recommendations both reflect visibility and player engagement). In that situation, individual coefficients should be interpreted cautiously: their signs and magnitudes can shift because the model is separating very similar signals.

Multicollinearity does not affect the interpretation of price, since *log_price* has *VIF* 1.12. Therefore, the negative coefficient of price can be interpreted as a stable partial relationship: higher prices are associated with slightly fewer owners, although the magnitude is small compared to visibility signals such as review volume.

After controlling for closely related visibility/quality variables, price has at most a small partial association with owners. This suggests that in the indie market, audience size is driven more by discoverability and perceived value (captured by reviews and engagement signals) than by price alone.

4.7 Conclusion

We grouped indie games into three success tiers using clustering, and the separation between tiers was driven most strongly by overall audience approval (*pct_pos_total*). The distribution is highly uneven: most indie games fall into the unsuccessful or somewhat successful tiers, while the highly successful tier is rare, reinforcing the idea that standout success is uncommon and risk is structurally high in the indie market.

Does genre matter?. Successful games are heavily concentrated in a few genre groups, dominated by Action and Casual, with the most common combination being {*action*, *casual*}. Many other frequent pairs include casual as the "bridge" (e.g.,tabletop/ simulation/ strategy + casual), suggesting that successful indies often mix a clear genre identity with accessible play patterns, while still adding depth through RPG/ strategy/ simulation-style progression.

Do mechanics matter?. The $genre- > mechanic$ rules show consistent associations: resource management is strongly linked to simulation/ strategy (and action hybrids), while exploration repeatedly complements action combinations (platformer, tabletop, RPG). This supports the idea that successful games often align mechanics with what players expect from that genre blend.

Do game characteristics matter?. The strongest recurring characteristic pattern is multiplayer design that combines Co-op + PvP, while VR-only appears but with low support, suggesting it's a niche path rather than a mainstream success driver.

Does pricing matter?. Only weakly compared to visibility and engagement signals. In the paid-only regression, price has a small negative association with owners, and its standardized effect is much smaller than review volume, which is by far the strongest correlate of audience size. Multicollinearity mainly affects review/ recommendation signals ($VIF \approx$ 6–7), but price is stable (low VIF) and still relatively minor.

Overall, these results are correlational, not causal—they don't prove that picking a genre or adding a mechanic causes success. However, they provide practical, evidence-based guidance to understand the market audience and reduce risk, helping us make more informed game-design decisions around genre direction, mechanic fit, feature scope, and pricing expectations.

5 Can a Single Game Influence Tag Adoption in Subsequent Releases?

This section studies whether the release of a highly successful game can influence the adoption of tags by games released after it. In particular, we analyze whether this influence appears as a change in the temporal trend of tag usage.

5.1 Tag Study Methodology

Instead of relying only on total tag counts, we focus on the temporal evolution of tag adoption. While overall popularity can be measured by the total number of games using a tag, this does not capture changes in adoption behavior over time.

We measure the monthly number of newly released games that include a given tag, both before and after the release of a target game. For each period, we fit a linear regression model to estimate the adoption trend. A significant difference between the pre-release and post-release slopes is interpreted as evidence of a change in adopting the tag.

5.1.1 Generalized Findings

The methodology described above can be generalized across all games and all tags in the dataset, allowing us to identify games that appear to have bigger influence on tag adoption patterns.

To focus on candidates most likely to have systemic impact, we restrict our analysis to games with an estimated ownership exceeding 2,000,000 users. These high-visibility titles form a subset of games plausibly capable of shaping market conventions and genre definitions.

Applying the trend-shift analysis to this subset, we find that approximately 96% of popular games in the dataset are associated with a statistically significant increase in the adoption rate of at least one tag following their release. This result suggests that tag influence is a common phenomenon among high-impact titles rather than an isolated occurrence.

5.2 Case Study: The Binding of Isaac and The Binding of Isaac: Rebirth (***Roguelike***)

The Binding of Isaac (2011) and its remake, *The Binding of Isaac: Rebirth* (2014), are considered founders of the roguelike genre.

We have applied the same trend change analysis to assess their influence on the adoption of the *Roguelike* tag. This trend change can also be clearly observed just by looking at the following figure showing the yearly frequency of games released under the tag "Roguelike"

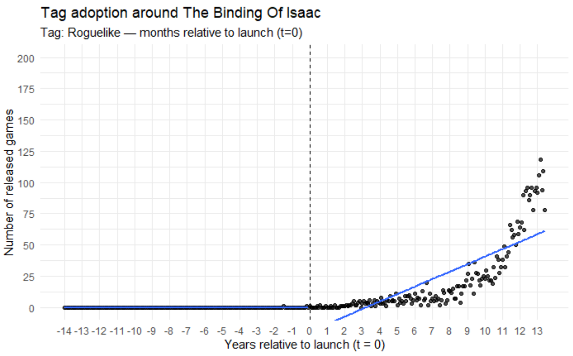


Figure 11: *The Binding of Isaac*

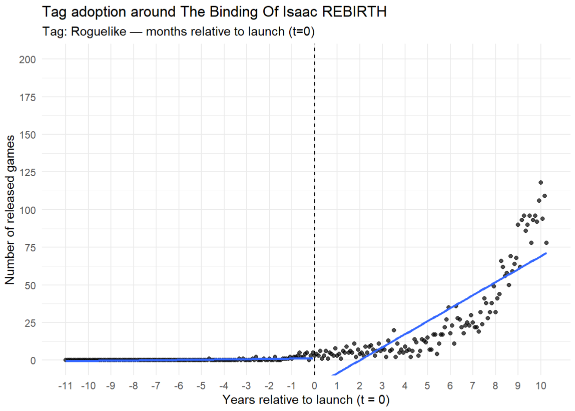


Figure 12: *The Binding of Isaac Rebirth*

Figure 13: Adoption of the *Roguelike* tag around the release of *The Binding of Isaac.*

Our findings indicate a sustained long-term increase in the usage of the *Roguelike* tag following the release of these titles, suggesting not only short-term influence but also lasting effects on genre classification practices within the platform.

5.2.1 Shared Influence and Non-Obvious Trend Behavior

Although both games show a big influence on tag adoption, visual inspection alone can be misleading. Both releases occur during a period where the tag was already growing, which could suggest that the observed effect is driven by broader market trends.

This highlights the limitation of relying only on visual analysis. The trend-change approach focuses on changes in the direction of adoption over time, allowing it to capture gradual or overlapping effects that are not immediately apparent in raw plots.

In addition, the analysis is constrained by an ownership threshold of 2,000,000 users. Games below this level are unlikely to generate enough visibility to affect platform-wide tagging behaviour. Given their strong sales performance, shared franchise, and close release dates, it is reasonable to treat both titles as a combined influential force within the *Roguelike* tag*Roguelike* tag

5.3 Case Study: Slay the Spire (***Roguelike Deckbuilder***)

Slay the Spire, released on January 23, 2019, is widely known as a defining title of the *Roguelike Deckbuilder* tag.

We analyze the frequency of games released with this tag before and after its launch, comparing the observed trend against the overall rate of game releases on Steam to control for platform-wide growth effects.

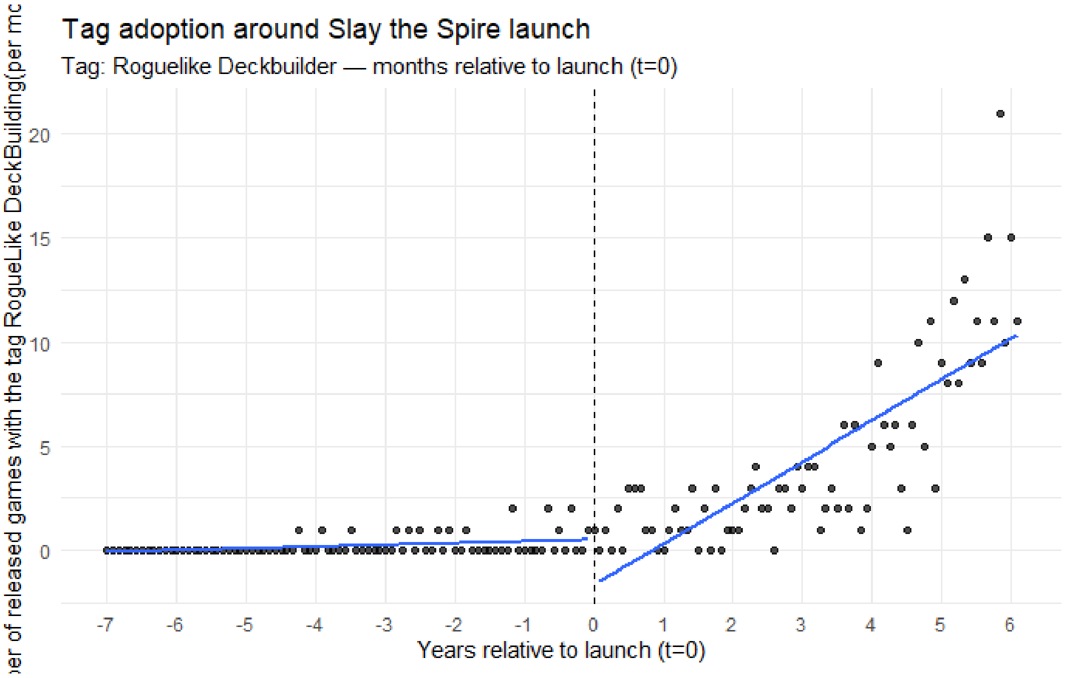


Figure 14: Adoption of the *Roguelike Deckbuilder* tag before and after the release of *Slay the Spire*.

The results show a significant increase in the post-release adoption rate of the *Roguelike Deckbuilder* tag, indicating that *Slay the Spire* played a major role in popularizing and formalizing this genre label on Steam. As discussed previously, this result indicates influence rather than direct causation.

The scale of the influence differs from broader genres: while many popular tags reach hundreds of new games per year, the *Roguelike Deckbuilder* tag remains more specialized.

5.4 General examples: Terraria (*Open World Survival Craft*) and Among Us (*Social Deduction*)

To further illustrate how highly successful games can influence tag adoption, we analyze two additional examples: *Terraria* and *Among Us*. Both titles achieved exceptional popularity and are widely associated with the formalization of their respective genre tags.

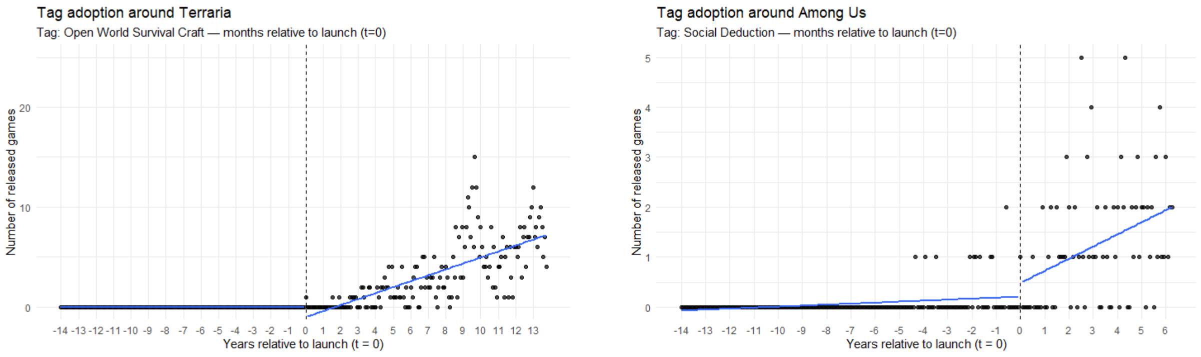


Figure 15: *Terraria*

Figure 16: *Among Us*

Figure 17: Adoption of the *Roguelike* tag around the release of *The Binding of Isaac*.

Terraria, released in 2011, is considered a defining title of the *Open World Survival Craft* genre. Figure 15 shows the adoption of this tag before and after the game's release. Prior to release, the tag appears only sporadically, with no clear growth pattern. Following the release of *Terraria*, there is a clear and sustained increase in the number of newly released games using this tag. The positive post-release trend suggests a long-term influence, indicating that *Terraria* contributed to establishing this genre as a recognizable and reusable category within the Steam ecosystem.

Among Us, released in 2018, is widely regarded as the game that popularized the *Social Deduction* genre. Figure 16 presents the adoption trend of this tag around its release. Before launch, the tag is rarely used and shows little variation. After release, a noticeable increase in adoption occurs, followed by a steady upward trend. Although the absolute number of games remains lower than for broader genres, the clear change in slope indicates that *Among Us* played a key role in defining and promoting this genre label.

Taken together, these examples complement the previous case studies. While *Terraria* shows a broad and sustained influence on a large genre category, *Among Us* demonstrates how a single highly visible title can establish a more specialized genre. In both cases, the release of a landmark game coincides with a persistent change in tag adoption dynamics, reinforcing the idea that successful games can shape how future titles are categorized on Steam.

5.5 Conclusion

The analysis in this section shows that changes in tag adoption trends can be detected following the release of highly successful games. By comparing pre-release and post-release adoption rates, we observe that landmark titles are frequently associated with sustained increases in the use of specific genre and mechanic tags.

Both the generalized results and the individual case studies indicate that this influence is not limited to short-term spikes. Instead, successful games often coincide with longer-term shifts in how tags are adopted by subsequent releases.

These findings do not establish causality. Rather, they provide quantitative and visual evidence that highly visible games can shape descriptive conventions and genre labels within the Steam ecosystem. This influence may operate at different scales, affecting either broad genre categories or more specialised niche tags (like *RogueLike DeckBuilder* or *Social Deduction*, depending on the scope and visibility of the title.

6 General conclusions

In this project, we analyzed a large Steam dataset (94,948 games, 47 variables) to study two related questions: how genres evolve over time on the platform, and what patterns are common among successful indie games. To make the data suitable for analysis, we first performed a preprocessing pipeline that reduced dimensionality, removed noisy observations (like playtests and invalid entries), handled duplicates, formatted key variables (dates and owner ranges), and built a grouped genre taxonomy to reduce inconsistencies in Steam's original labels.

From the exploratory analysis of genre lifecycles, we observed that genre popularity is better studied through market share rather than raw counts, since the platform has grown significantly over time. Using market share trends, we found that some genres gained relative importance (notably Casual, and also Action and Roguelike), while others remained niche or decreased in share (e.g., Music, Tabletop, and Shooter). We also observed that player engagement is highly skewed in all genres: mean playtime is much higher than median playtime, indicating that a small group of highly engaged players strongly influences average values. Some genres, especially Music, show particularly strong long-tail behavior, suggesting that niche genres can still generate very high engagement among a small audience.

To study success in the indie market, where success is not explicitly labeled, we built an operational proxy using clustering over engagement and market signals. This approach allowed us to isolate higher impact games and focus pattern mining on a meaningful subset. Within successful indie games, we found recurring genre combinations (especially those involving Action and Casual), and we observed that Casual often acts as a bridge in hybrid games (e.g., Strategy + Casual, Simulation + Casual). We also identified strong associations between genre mixes and mechanics: resource management appears strongly linked to Strategy/Simulation hybrids, while exploration frequently complements Action-based combinations. In addition, some experience characteristics repeat among successful games, such as multiplayer modes combining Co-op and PvP, while niche formats like VR-only appear less frequently. Finally, the pricing analysis suggests that price alone has a weak relationship with owners compared to visibility and engagement signals such as review volume.

Lastly, we studied whether single landmark games can influence the adoption of tags in later releases. By comparing pre-release and post-release adoption trends, we observed that highly visible titles are often followed by sustained increases in the use of specific tags. Case studies such as *The Binding of Isaac*, *Slay the Spire*, *Terraria*, and *Among Us* show trend shifts consistent with the idea that influential games can help formalize genre labels and shape how future games are described on Steam. While these effects are correlational and do not prove causality, they provide consistent evidence that successful games can have broader ecosystem impact beyond their own audience.

Overall, the results of this project highlight two key takeaways: Steam genres are dynamic and shaped by external influences and player behaviour, and successful indie games tend to cluster around recurring genre-mechanic patterns rather than a single formula. The methods used (preprocessing, exploratory analysis, clustering, association rules, regression, and trend-shift analysis) provide a structured way to extract knowledge from large-scale market data, and to better understand the factors that repeatedly happen in high-impact indie games.