Steam Games Project

Samuel Escalante Gutierrez
Manuela Mayorga Rojas
Mariana Mera Gutierrez
Universidad Autónoma de Occidente
March, 2024

Introduction

The video game industry is undergoing incredible growth, with intense competition among developers to create the next big success. In this scenario, Steam is recognized as the dominant digital platform for PC games, boasting approximately 50,000 available games and over 30 million active players.

Launching a game on Steam presents a significant challenge due to the fierce competition in this market. With thousands of games available, capturing players' attention is crucial. However, having accurate information and effective tools can make a difference in the chances of success. This is where the chosen dataset on Steam comes into play, providing valuable information about the Steam market, such as genre popularity and player behavior.

This project is designed to benefit anyone interested in releasing a game on Steam, including experienced game developers and publishers. With the right information and tools, you can significantly increase your chances of success on this highly competitive platform.

Objectives

The objective of this project is to conduct a detailed analysis of the Steam dataset to create visualizations that allow us to understand current trends in the gaming industry. Our mission is to provide game developers and publishers with essential tools and information to improve their chances of success on the Steam platform.

To start, we will focus on developing data-driven analyses of STEAM to identify dominant trends in the gaming industry, highlighting key publishers and developers. Subsequently, we will explore trends in genres and types of games, analyzing which ones are the most popular.

In addition to providing information about key features of successful games on STEAM, we aim to offer valuable data for new developers to emulate successful strategies. Ultimately, we seek to provide game developers with a comprehensive understanding of current industry trends, addressing aspects such as genre, pricing, and player preference for single-player or multiplayer games.

Therefore, our goal is to equip gaming industry professionals with the necessary information to make informed decisions, fostering success in a highly competitive environment like the Steam platform.

For this project, we will be using the following tools:

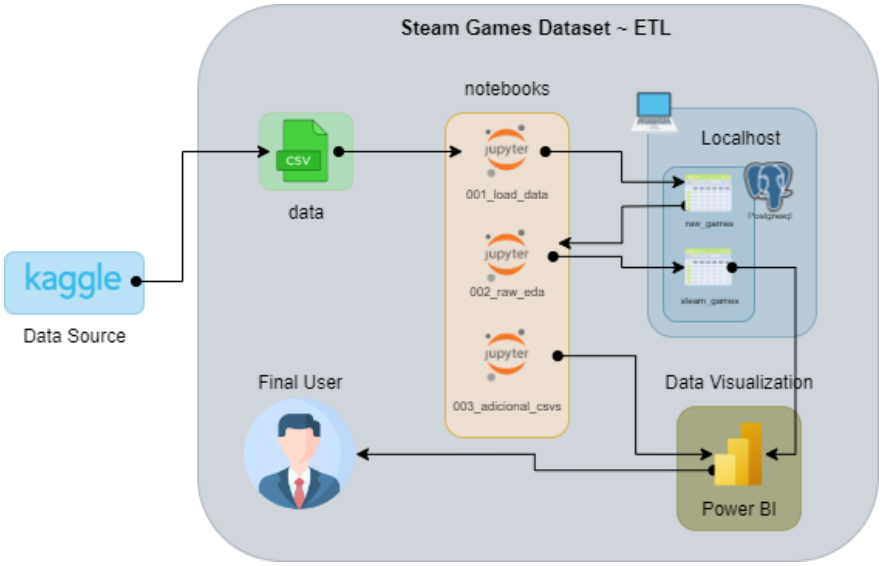
Python: Python is a versatile and popular programming language that is ideal for data analysis. It is easy to learn and use, with a wide range of libraries available for data analysis tasks.

SQLAlchemy: The cornerstone of our project will be SQLAlchemy, a powerful Object-Relational Mapping (ORM) library for Python. This framework will facilitate interaction with the PostgreSQL database, enabling efficient data management and simplifying queries through an object-oriented interface.

PostgreSQL: PostgreSQL is an open-source SQL database that is robust and scalable. It is a good choice for projects requiring a reliable and high-performance database.

Jupyter Notebooks: Jupyter Notebooks is an interactive environment for executing Python code and visualizing results. It is an ideal tool for exploring data and developing analysis prototypes.

Power BI: Allows the creation of interactive panels, reports and dashboards, facilitating the presentation and understanding of results through dynamic and customizable visualizations and connects directly to data sources, automatically updating reports as data evolves.



Workflow

Project Steps:

1. Data connection and upload
2. Data cleaning
3. EDA
4. Upload finall dataset into database
5. Additional CSV
6. Power BI Graphics
7. Conclusions

1. Database Connection

Environment Configuration: Before starting the data loading process, it is necessary to ensure that the environment is properly configured. A '.env' file must be present in the working directory containing the 'WORK_DIR variable', specifying the path of the working directory.

1.1. Database Connection

To establish a connection with the PostgreSQL database, SQLAlchemy, a Python library for managing relational databases, is used. A SQLAlchemy session is created to interact with the database, and it is verified whether the 'raw_games' table already exists in the database. If the table exists, it is dropped and recreated. If it doesn't exist, it is created directly.

1.2. Loading Data from a CSV File

Data loading is performed from a CSV file named 'games.csv', containing relevant information about Steam games. To process and load the data into the database, the Python Pandas library is utilized. Before loading the data, a transformation is applied to capitalize the first letter of each word in the column names. The data is then loaded into the 'raw_games' table of the PostgreSQL database using Pandas 'to_sql()' method.

2. Data Cleaning

In our study on the analysis of video games on the Steam platform, data cleaning is presented as a fundamental step to ensure the validity and reliability of our research. This process allows us to address challenges such as missing values and inconsistencies, preparing our data for detailed and meaningful analysis.

The provided table reflects the complexity and richness of data related to video games on the Steam platform. It contains a wide variety of attributes, from the game ID to the estimated number of owners, Metacritic scores, average playtime, and more. However, this wealth of data also exposes common challenges in data cleaning, such as the presence of missing values and duplicate entries. Therefore, data cleaning becomes an essential step to address these imperfections and ensure the integrity of our analyses.

2.1 Column Removal

Removing columns in a dataset is carried out with the aim of improving the quality and relevance of the data for analysis. In this case, the removed columns are justified for several reasons:

Null Values

Some of the removed columns contain a significant amount of null values. In statistical analysis, missing values can distort results or make proper interpretation of the data challenging. By eliminating these columns with a large number of null values, we ensure that our dataset is more complete and reliable for subsequent analysis.

Lack of Value for Analysis

Other removed columns, while they may contain valid information, are not relevant to the specific objectives of our analysis. For example, columns related to image links, websites, promotional material, and detailed descriptions may not directly contribute to our main analysis on factors influencing the success of video games on the Steam platform. By removing these columns, we simplify our dataset and focus on variables that are more pertinent and significant for our study.

Name	Data Type
AppID	0
Name	0
ReleaseDate	0
EstimatedOwners	0
PeakCCU	0
RequiredAge	0
Price	0
DLCCount	0
SupportedLanguages	0
FullAudioLanguages	0
Windows	0
Mac	0
Linux	0
MetacriticScore	0
UserScore	0
Positive	0
Negative	0
Achievements	0
Recommendations	0
AveragePlaytimeForever	0
AveragePlaytimeTwoWeeks	0
MedianPlaytimeForever	0
MedianPlaytimeTwoWeeks	0
Developers	3581
Publishers	3861
Categories	4595
Genres	3553
Tags	21094

2.2. Imputation of Missing Data

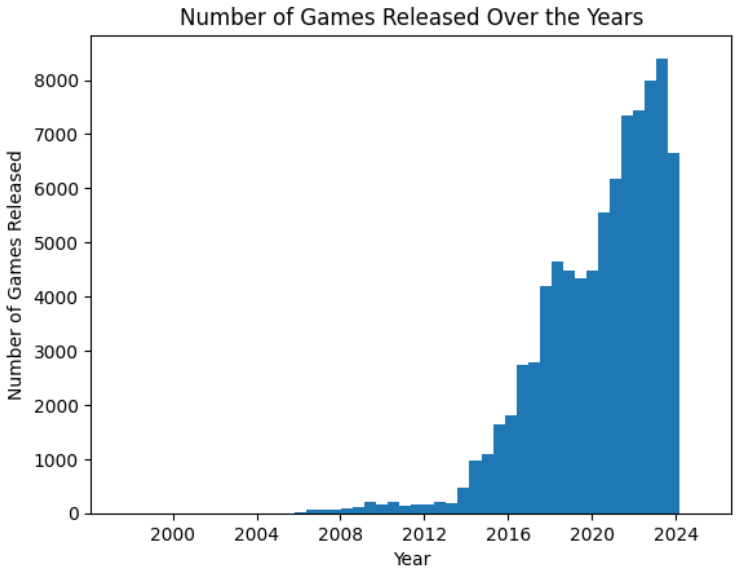
Data imputation is a crucial process in handling missing values in a dataset. In this case, imputing data for the 'Developers,' 'Publishers,' 'Categories,' 'Genres,' and 'Tags' columns is justified for several reasons.

These columns are descriptive in nature and provide information about key aspects of video games, such as who developed and published them, the categories they belong to, their genres, and associated tags. Since the absence of this information can be detrimental to the analysis, especially if exploring relationships between these variables and other attributes, data imputation helps preserve the integrity of the dataset and ensures that these features are not overlooked during analysis.

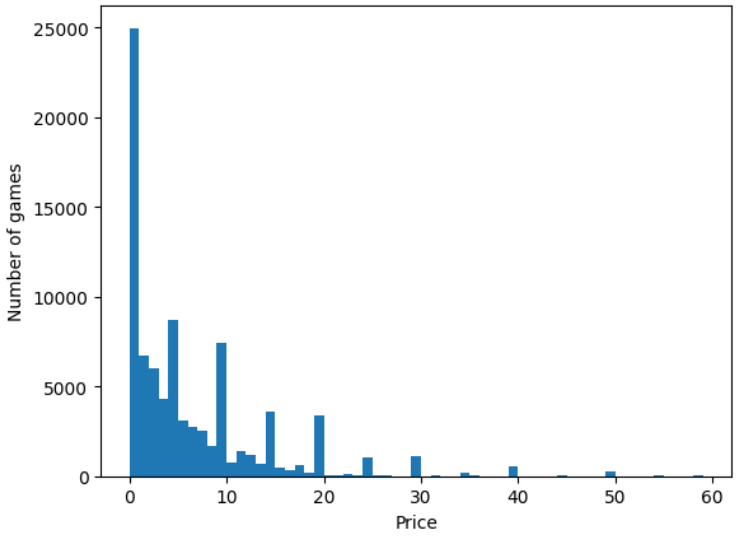
The choice to replace missing values with 'Others' in these columns is based on the premise that, although specific data may not be available, we can still infer that games have developers, publishers, categories, genres, and associated tags, even if not explicitly recorded in the dataset. By assigning the value 'Others', we retain the structural information of the dataset and avoid significant distortions in subsequent analyses that may arise from removing records with missing values in these columns.

3. EDA (Exploratory Data Analysis)

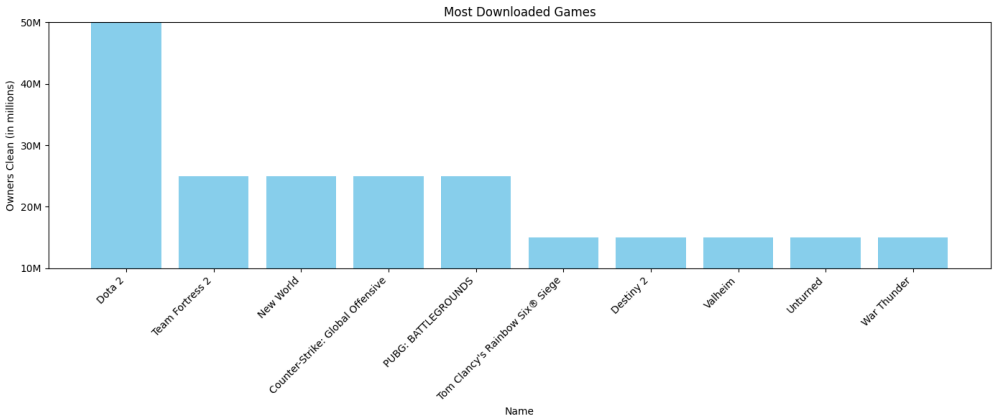
Exploratory Data Analysis (EDA) is a fundamental stage in the investigation of any digital environment, and the Steam platform, as a leader in the video game market, is no exception. This process allows us to delve into the vast amount of data available on Steam to better understand the trends, behaviors, and patterns that influence the reception of video games by users.



Analysis of the bar chart reveals steady growth from 2016 until reaching a peak in 2023, reaching 8000. This sustained growth suggests a positive trend on the Steam platform. In addition, an evolution in the number of games released over the years is observed, indicating that the time required to publish a game has decreased.



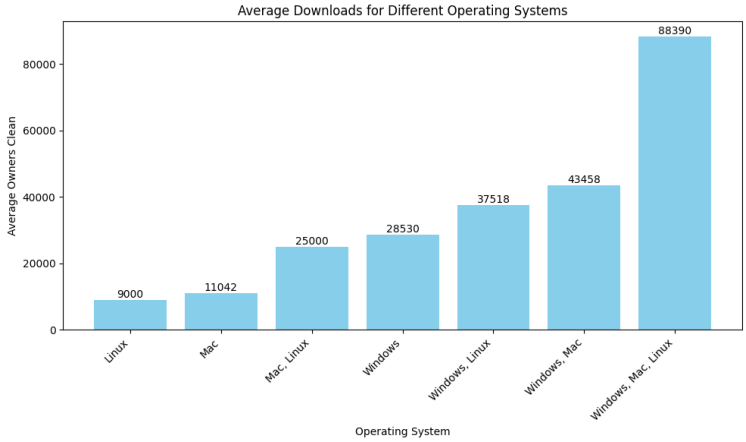
Analysis of the price distribution reveals a noteworthy phenomenon: a large number of free games on the Steam platform. The peak at the lower end of the graph, with approximately 25,000 games priced at $0, indicates significant diversity of available free games. The presence of free games enhances the user experience by providing them with entertainment options at no initial cost.



The analysis of the 'Most Downloaded Games' graph reveals interesting patterns regarding the popularity of certain games on the Steam platform. A notable peak of 50 million downloads stands out for the game Dota 2, indicating its dominant position as the most downloaded game on Steam. Furthermore, a consistent trend is observed in other popular games such as Team Fortress 2, New World Countries Strike, and PUBG, all boasting around 25 million downloads. This distribution of downloads suggests the presence of an active and engaged player base on Steam, with a wide variety of games attracting millions of users.

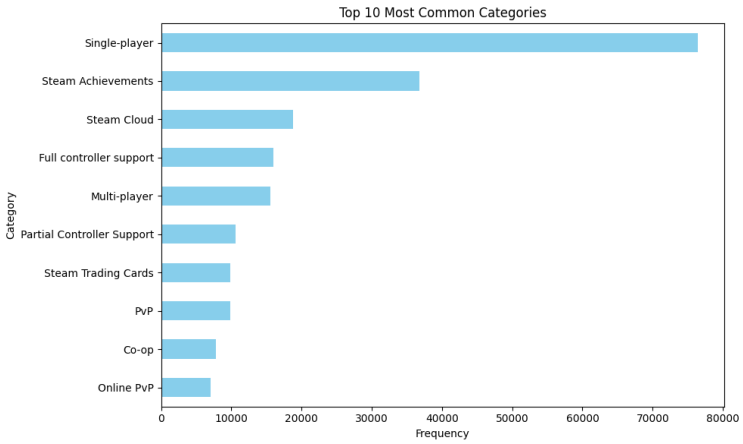
Column OS

The new 'OS' column is created to display in a clear and concise manner the operating systems compatible with each game in the Steam platform dataset. This information is essential for users who want to quickly determine if a game is compatible with their operating system. The column is generated through a function that evaluates the compatibility options of each game and presents them in a readable format within a single column.

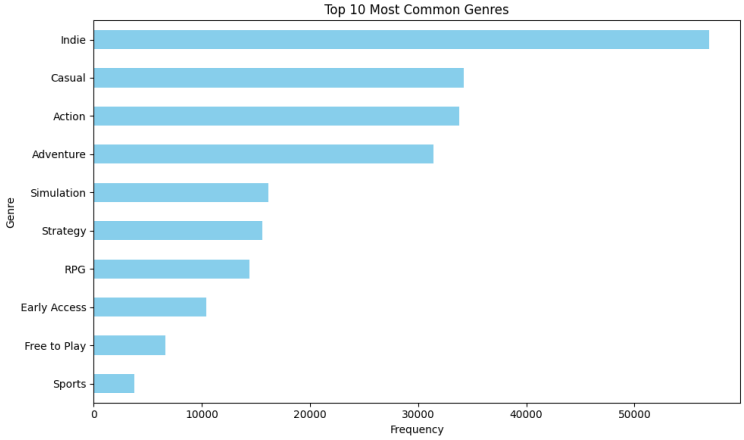


The analysis of the relationship between operating systems (OS) and the average number of owners (OwnersClean) of games on the Steam platform reveals interesting patterns. It is observed that games exclusively compatible with Windows have the highest average number of owners, around 28,530 on average. This suggests that the majority of popular games on Steam are primarily designed for the Windows platform.

On the other hand, games that are compatible with multiple operating systems tend to have a higher average number of owners compared to games exclusive to a single operating system. For example, games compatible with Windows, Mac, and Linux have the highest average number of owners, approximately 88,390. This suggests that versatility in terms of compatibility with different operating systems can increase the potential user base of a game.



The analysis of the ten most frequent categories in Steam platform games reveals significant trends in player preferences. The high demand for single-player experiences stands out, emphasizing the importance of Steam achievements and features such as Steam Cloud and full controller support. Furthermore, the popularity of multiplayer and cooperative options indicates a continued interest in social interaction and online competition. Features like Steam trading cards add an additional dimension to gameplay and community engagement. Overall, these trends suggest a diverse range of player interests, combining solo gaming experiences with social interaction and competitive elements.



The chart shows the 10 most common video game genres, based on the frequency with which they appear in a data sample. The indie (games created by small teams or independent developers.) and casual (simple and easy-to-learn games, generally aimed at casual gamers.) genres are the most common, indicating that there is a high demand for games that are accessible and easy to play. On the other hand, action, adventure and simulation genres are also very popular. These genres offer a wide range of gaming experiences, from first-person shooters to city-building games. The sports genre is the least common, which could be because there is less demand for sports games compared to other genres. Overall, the chart provides a good overview of the most popular video game genres. This information can be useful for game developers who are trying to decide what type of game to create.

Column Supported Languages

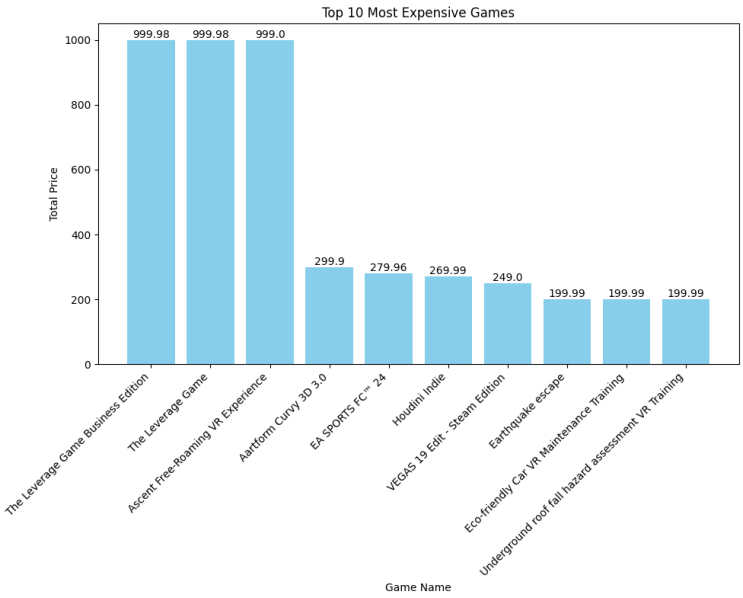
The original data in the 'Supported Languages' column consisted of language lists represented as text strings enclosed in square brackets and single quotes. However, some values included unwanted characters such as HTML tags. These characters were removed during the data cleaning process using string manipulation

methods in Python. Additionally, square brackets and single quotes were removed, leaving only the names of the languages.

Language	Frequency
English	78060
Simplified Chinese	19306
German	18652
French	18075
Russian	17383
⋮	⋮
Hungarian,Polish	1
Japanese (all with full audio support)	1
English,German,Spanish - Spain,lang_français	1
Korean	1
Traditional Chinese (text only)	1

The analysis of Steam dataset reveals that English is the most predominant language on the platform, with 78,060 mentions. It is followed by Simplified Chinese with 19,306 mentions, and then German with 18,652 mentions. French and Russian also have a significant presence, with 18,075 and 17,383 mentions respectively. This pattern reflects the linguistic diversity on the platform, indicating an international audience and the adaptation of games to meet the language preferences of Steam users.

Top 10 Most Expensive Games



The bar chart displays the top 10 most expensive games, based on their total price. The first 3 positions are occupied by different versions of "The Leverage Game," all priced at around $999. This suggests that it is a highly comprehensive or premium gaming experience, or possibly one targeting a specific audience willing to pay a significant price.

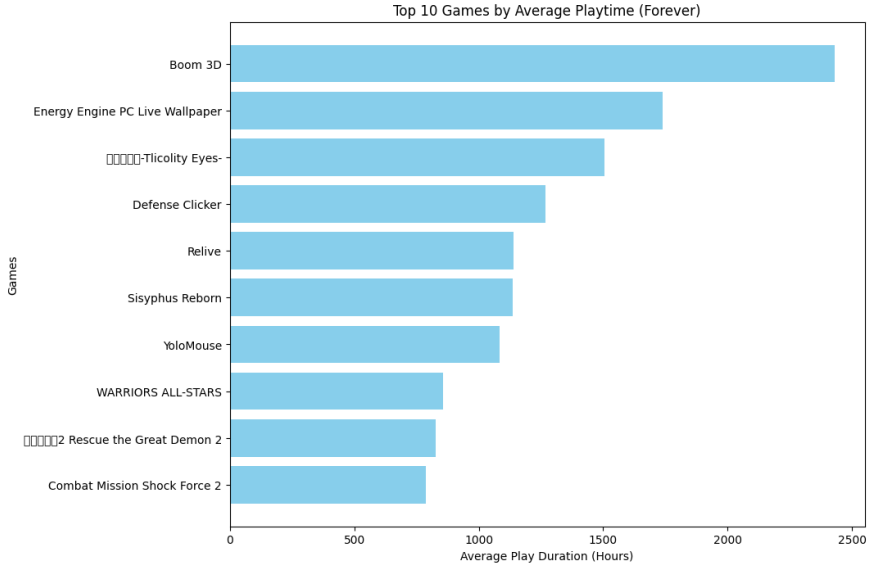
There is a drastic decrease in price after the top 3 games. This indicates a clear distinction between the most expensive game and the rest. It could imply that the top 3 are considered "luxury" or premium experiences, while the others fall into a more standard pricing category.

Free or Paid

FreeOrPaid	Name
free	16456
paid	68641

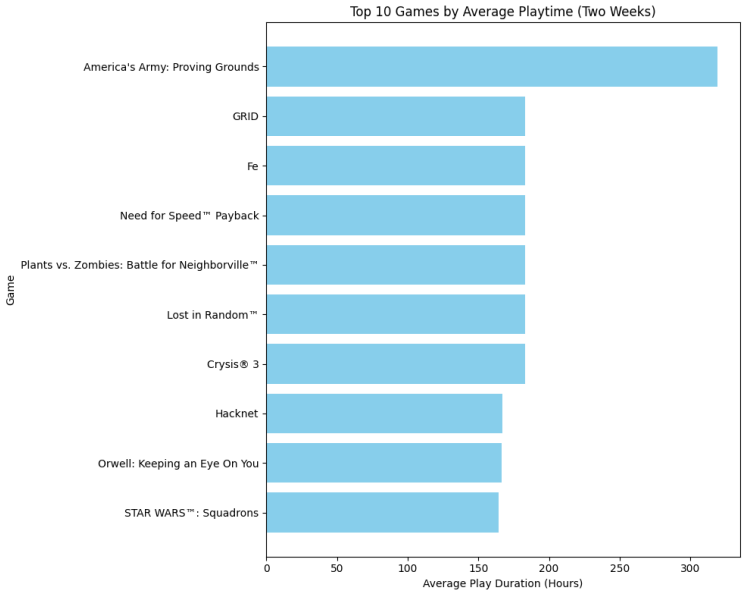
The analysis of the distribution of games based on their condition of being free or paid on the Steam platform reveals a prominent trend: a large proportion of games are paid, with 68,641 games, while 16,456 games are free. This suggests that although there is a variety of free games available, the majority of games on Steam require payment to access them.

Top 10 Games by Average Playtime (Forever)



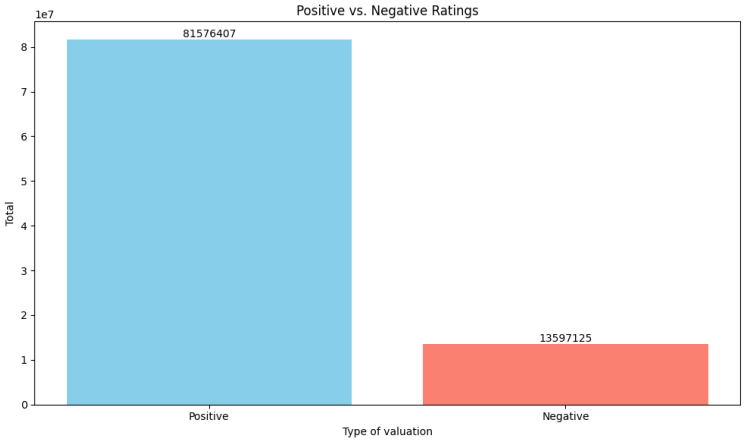
Analysis of the bar chart "Top 10 Games by Average Playtime (Forever)" reveals interesting patterns in terms of average playtime for the most popular games on the Steam platform. A notable peak is observed for the game Boom 3D, with almost 2500 hours of average playtime, followed by Energy energine PC Live wallpaper with around 1800 hours. However, as we move down the list, the average play time gradually decreases, with most of the remaining games below 1500 hours and some even below 1000 hours. This pattern suggests that while some games may generate exceptionally high player engagement, most games have more moderate play times in comparison.

Top 10 Games by Average Playtime (Two Weeks)



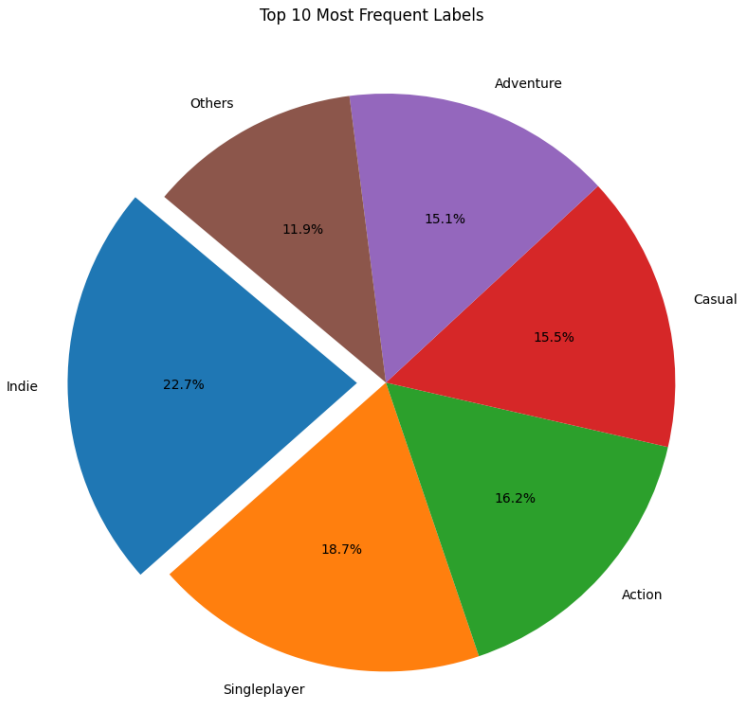
The chart shows the 10 games with the highest average playing time over two weeks in January. The games with the highest average play time are mostly action, adventure and shooter games. This indicates that these genres are the most popular among gamers who are willing to spend a lot of time on a single game. There are two racing games on the list, Need for Speed™ Payback and GRID. This suggests that racing games can also be very appealing to gamers looking for a challenging and rewarding gaming experience.

Positive vs. Negative Ratings



The graph shows the distribution of positive and negative game ratings. Most of the games have a positive rating, this indicates that, in general, players are satisfied with the quality of the games that are available and there are a small number of games with a negative rating, this indicates that there are some games that are not so popular among players.

Top 10 Most Frequent Labels



Analysis of the Steam tag graph provides insight into the distribution of games in the dataset. Indie, is the tag with the highest percentage (22.7%), indicating that a large portion of the games in the dataset are indie, followed by singleplayer as the second most popular tag (18.7%), suggesting that most of the games are single-player experiences. Action, Casual and Adventure: follow in popularity (16.2%, 15.5% and 15.1% respectively), showing a variety of genres among games, and finally the Other category groups games with no tags or unassigned tags, and represents 11.9% of the total.

4. Upload finall dataset into database

This part is designed to facilitate the loading of a previously cleaned and analyzed dataset into a database. It ensures that the table structure in the database is up to date and aligned with any changes made to the data during the cleansing and exploratory analysis processes.

4.1. Verification of the Existence of the 'steam_games' Table

Before performing the data load, the script checks if the table 'steam_games' already exists in the database. If the table exists, it proceeds to delete and recreate it. This is done to ensure that the table structure is updated according to the changes made to the dataset.

4.2. Creation or Recreation of the table 'steam_games'.

In case the table 'steam_games' does not exist, it is created directly. If the table already exists, it is deleted and recreated to ensure that any changes to the data structure are correctly reflected in the database.

4.3. Loading Data into the 'steam_games' Table

Once the table is ready, the script loads the cleaned and parsed data set into the 'steam_games' table in the database. The 'to_sql' method of the Pandas library is used to perform this load, and it is specified that, in

case the table already exists, it is replaced with the new data.

4.4. Error Management

At each stage of the process, error handling is implemented to handle possible problems that may arise during table creation or data loading. Any errors detected are printed on the console to facilitate problem identification and correction.

Finally, the database is logged off to maintain proper management of resources and connections.

5. Additional CSV

During the Exploratory Data Analysis (EDA) and the creation of the Power BI dashboard, we encountered a common challenge in our data: columns containing grouped records, such as lists separated by commas. We will employ data transformation and cleaning techniques to break down these records into individual columns, enabling us to explore in greater detail the distribution and frequency of elements within these lists.

5.1. Supported Languages

In this section, our focus is on analyzing the languages supported by games in the Steam dataset. We split the language lists into individual records and calculate the frequency of each language. The result is stored in a new CSV file named 'SupportedLanguages.csv', containing details on the frequency of each language.

5.2. Categories

In this segment, we delve into the categories assigned to games in the Steam dataset. Similar to the language analysis, we split the category lists into individual records and create a CSV file named 'Categories.csv' with detailed information about the distribution of categories.

5.3. Genres

This part concentrates on exploring the genres assigned to games in the Steam dataset. Similar to the previous analyses, we break down the genre lists into individual records and generate a CSV file named 'Genres.csv' with details on the distribution of genres.

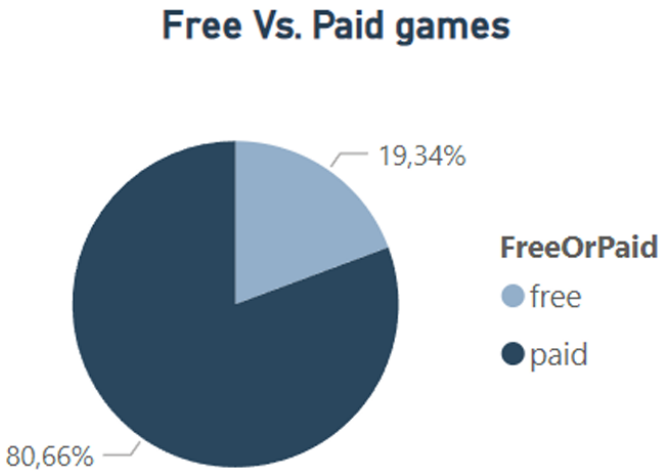
5.4. Tags

Lastly, we explore the tags assigned to games in the Steam dataset. A process similar to the language analysis is applied, splitting the tag lists into individual records. The result is stored in a CSV file named 'Tags.csv', providing detailed information on the distribution of tags associated with games on the Steam platform.

This additional analysis provides a deeper understanding of the linguistic diversity, categories, genres, and tags present in the dataset, allowing us to extract valuable insights for future research and analysis.

6. Power BI Graphics

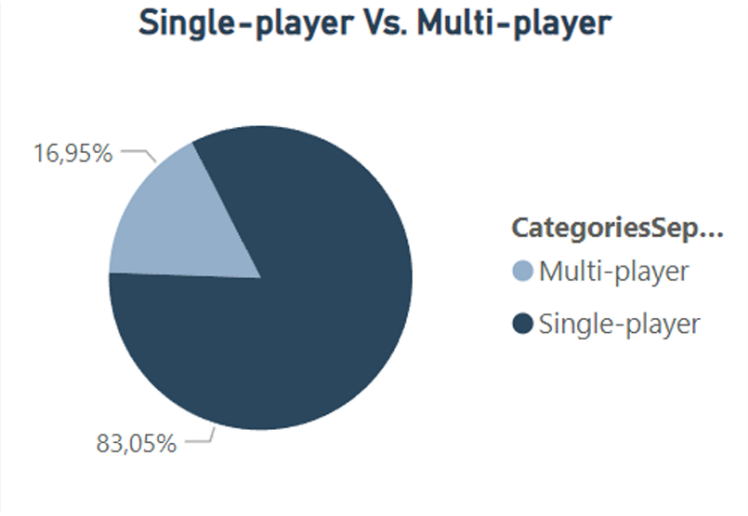
Free vs. Paid game pie chart



The majority of games, represented by 80% of the total, belong to the free category, with a total of 16,456 titles. On the other hand, the paid games category constitutes the remaining 20%, with a total of 68,641 games.

This analysis suggests a marked trend towards free-to-play in the video game industry, possibly driven by business models based on advertising, in-app purchases or other strategies to attract a wide audience. However, the significant presence of paid games highlights the coexistence of varied business models, indicating that consumers still value quality and gaming experience enough to invest in premium titles

Singler-Player vs Multi-Player game pie chart



While multiplayer remains relevant, especially in the realm of online gaming, the preference for solo experiences suggests a continued demand for engaging stories and the ability to explore virtual worlds alone.

These findings are critical for game developers and designers, as they indicate the need to balance the supply of titles to satisfy both players seeking social experiences and those who value intimate connection to the

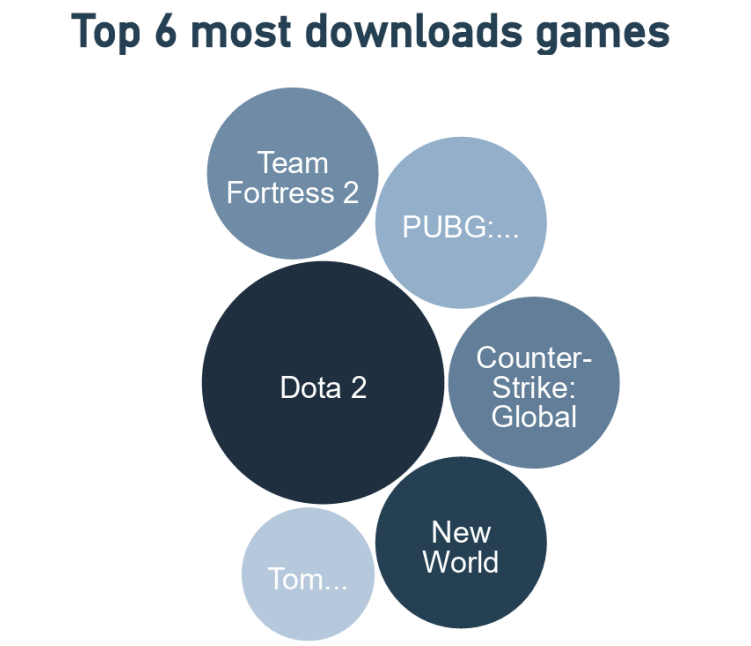
plot and gameplay, and why not integrate both experiences in the same game. Diversification of game modes remains essential to maintain vitality and relevance in the video game industry, offering diverse options for an increasingly heterogeneous audience.

Cloud graph with top 15 most used language

The predominance of languages such as English and Simplified Chinese suggests a focus on key global markets, while the presence of languages such as Spanish (Spain and Latin America) highlights attention to Spanish-speaking audiences in different regions of the world.

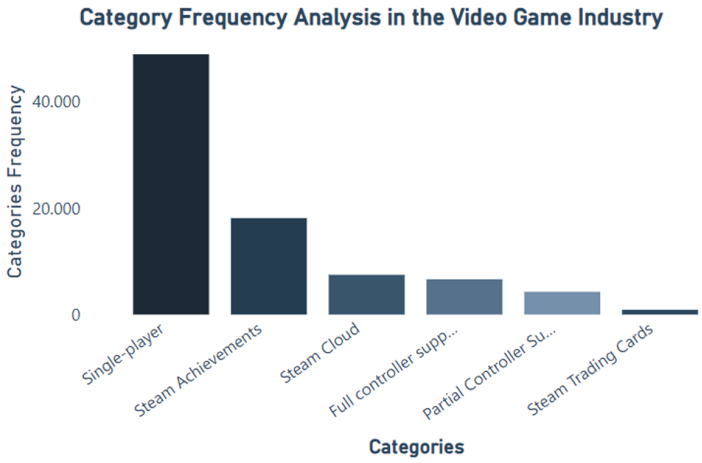
This analysis underscores the need for game developers and creators to consider localization and linguistic adaptation as essential elements in reaching and engaging a diverse audience. The inclusion of a variety of languages in games not only expands global reach, but also enhances the user experience by providing linguistic options that align with the cultural diversity of players.

Bubble chart of the Most Downloads Games



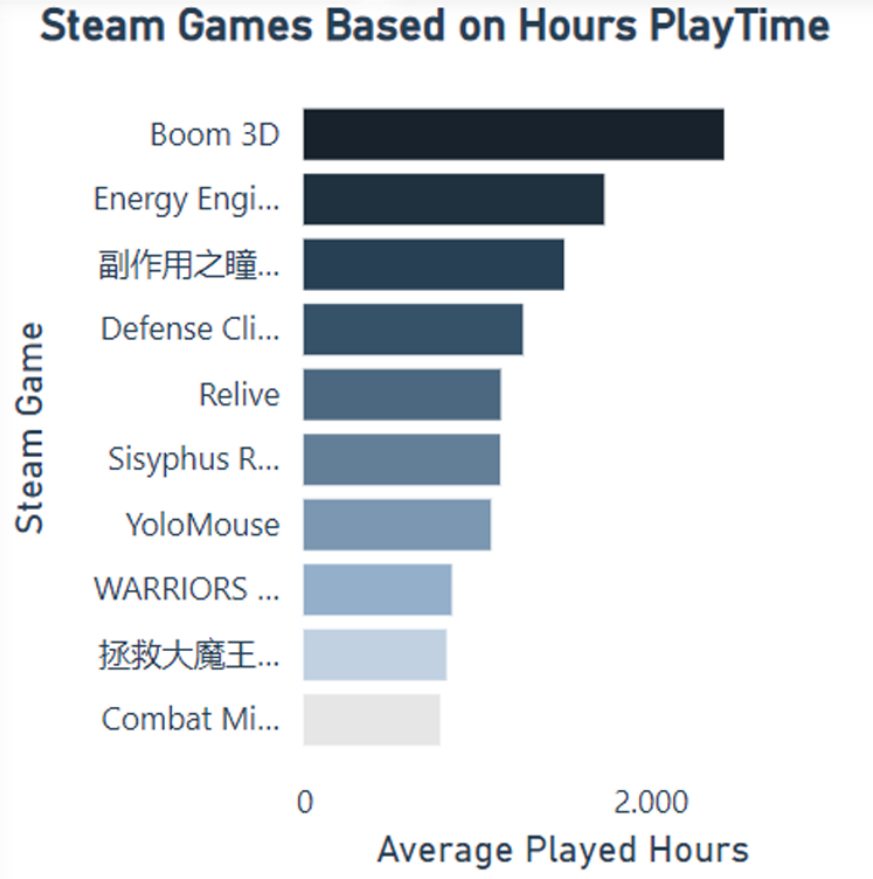
Overall, this bubble chart analysis highlights the diversity in popularity and features of the selected games. Dota 2 and Team Fortress 2 may represent stability and a loyal fan base, while PUBG: CS GO, New World and Tom Clancy's Rainbow Six could be leveraging novelty and innovation to attract new players. This visual approach provides a snapshot view of each game's relative position and relevance, allowing for a deeper understanding of their impact on the video game landscape.

Bar graph of Category Frequency Analysis in the Video Game Industry



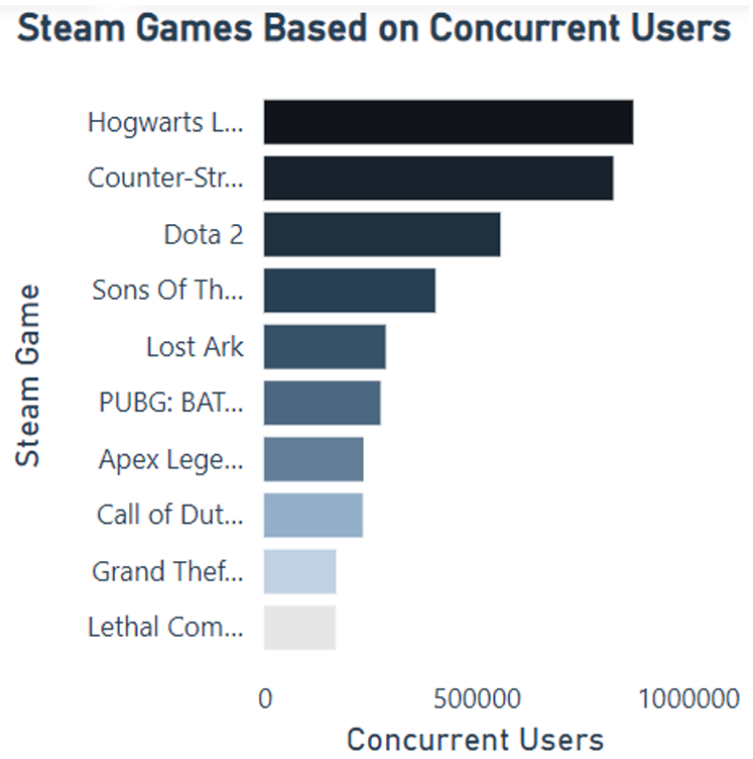
These trends indicate the diversity of preferences among gamers, from those who value the individual experience to those who seek social interactions and recognition. For the video game industry, this information is essential for strategic decision-making in game design and development, ensuring the creation of experiences that align with the changing expectations of the gaming community.

Horizontal bar chart of Steam Games based on Hours Playtime



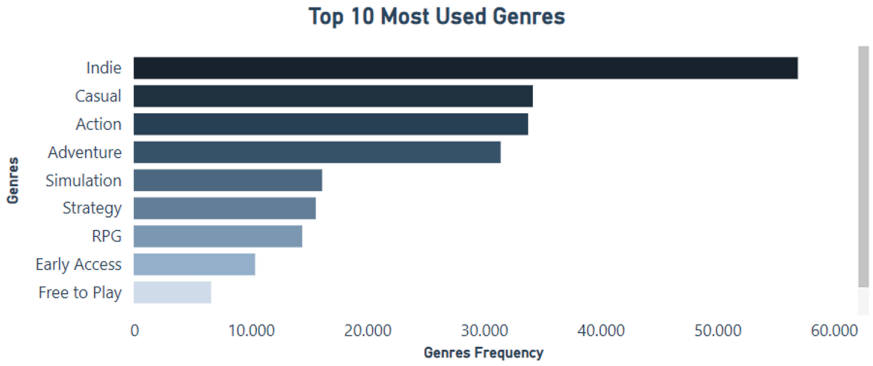
Through this visual representation, some titles that stand out significantly in terms of average hours played are highlighted, revealing that the aforementioned games have managed to not only capture the attention of gamers, but also maintain long-term engagement. This analysis provides valuable information for developers and content creators, highlighting the importance of creating deep and immersive gaming experiences that hold the attention and interest of the gaming community.

Horizontal bar chart of Steam Games based on Concurrent Users



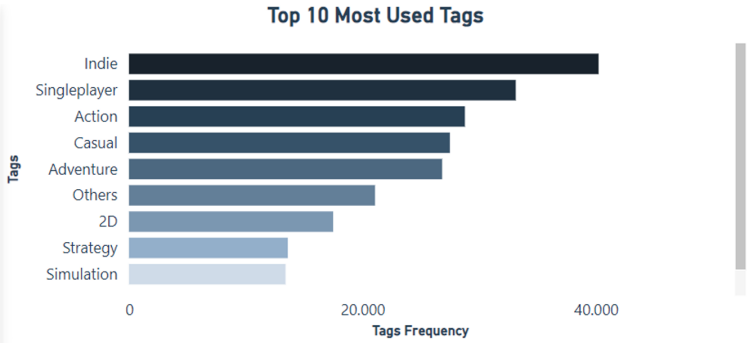
This analysis highlights the vitality of certain games that have managed to attract a large number of simultaneous players, the importance of the number of concurrent users as a key indicator of popularity and the impact of games on the gaming community. These titles have not only managed to attract a mass audience, but have also maintained their appeal, indicating quality and continued relevance in the dynamic video game market.

Horizontal bar chart of the 10 most used Genres



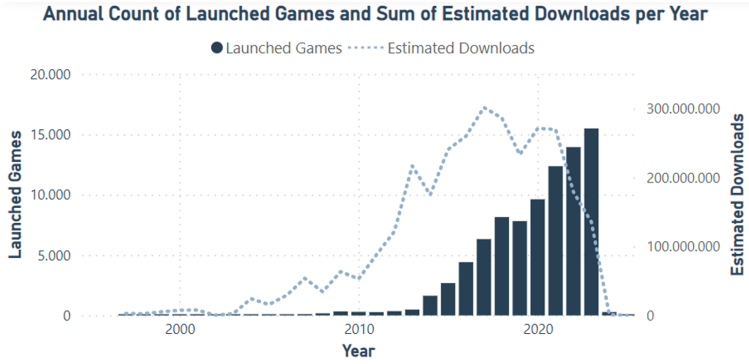
These results are crucial to understanding the diversity and distribution of genres that characterize the current video game landscape. This analysis suggests that the video game industry is marked by a wide variety of genres that cater to the diverse preferences of gamers. The preeminence of genres such as Indie and Casual reflects an appreciation for accessible and creative gaming experiences, while the inclusion of more traditional genres ensures a balance between innovation and familiarity for consumers. This knowledge is essential for developers and content creators, as it allows them to strategically adapt their offerings to meet the changing demands of a diverse audience.

Horizontal bar chart of the 10 most used Tags



This analysis provides a deep understanding of the preferences and values that define today's video game industry. Developers can use this information to adapt their design and development strategies, ensuring the creation of games that resonate with changing audience expectations and desires. The diversity of tags reflects a dynamic and constantly evolving industry, where innovation and adaptability are key elements for continued success.

Bar and line graph on Annual Count of Launched Games and sum of Estimated Downloads per year



Several trends and patterns are highlighted in this analysis, providing valuable information on market growth and dynamics.

Sustained Growth Trend: From the early 2000s to the present, there is a steady growth trend in both the number of games released and the estimated number of downloads. This indication suggests continued robust growth in the industry, supported by increasing demand for new content.

Boom in Releases after 2010: Starting in 2010, a significant increase in the number of games released per year is observed, coinciding with the boom in the video game industry. This period stands out for an exponential increase in production and diversification of titles.

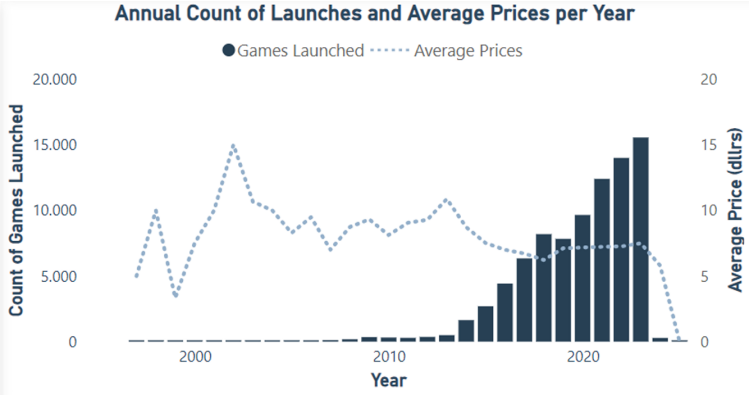
Stability in Downloads since 2018: Despite continued growth in the number of releases, download estimates appear to have reached a point of stabilization since 2018. This phenomenon may indicate a saturation in the supply of games or a change in consumer behavior.

Diversification in Game Offering: The variety of games released increases significantly over the years, indicating a diversification in genres, styles and platforms. This phenomenon may be a result of market expansion and the inclusion of new developers.

Impact of the Year 2020: The year 2020 is highlighted by a notable increase in the number of games released, possibly influenced by factors such as the COVID-19 pandemic, which led to increased interest and participation in the video game industry during periods of confinement.

Challenges in Year 2022: The year 2022 presents a significant decrease in the number of games released, which could indicate challenges or changes in market dynamics. This decline may require further exploration to understand the underlying causes. The saturation in downloads from 2018 and the decline in 2022 could signal a point of reflection for the industry in terms of release strategies and audience engagement. While game production remains high, developers may face challenges in standing out and attracting audiences in an increasingly competitive market.

Bar and line graph on Annual Count of Launches and Average Prices per year



Important patterns on the dynamics between game supply and pricing strategies in the industry.

Downward Trend in Average Prices: Over the years, there is a downward trend in the average price of games released. This decline may be indicative of more competitive and accessible pricing strategies, possibly in response to the growing demand for affordable games.

Price Stability After 2010: From 2010 to 2023, average prices remain relatively stable, indicating possible consolidation in pricing strategies and greater predictability for consumers.

Increase in Lower Priced Launches: The significant increase in the number of games launched from 2010 is associated with a decrease in average prices, suggesting a market strategy that prioritizes accessibility and breadth of supply.

Impact of Year 2024: The year 2024 is highlighted by a pronounced decrease in average price, possibly influenced by changes in industry pricing strategy or market-specific factors.

The analysis reflects the evolution in the video game industry's pricing strategy, characterized by a downward trend in average prices. This adaptation may be linked to the need to meet consumer expectations in a competitive market, as well as to the increase in the supply of games. The stability in prices after 2010 suggests greater maturity and consolidation in the industry. However, the decline in 2024 raises questions about market dynamics in that specific year.

6. Conclusions

The analysis of the Steam dataset has allowed us to obtain valuable information about current trends in the video game industry. Interesting patterns have been identified regarding game popularity, the most common genres, features most valued by players, and the most commonly used languages.

The analysis of prices and releases highlights an industry-wide strategy to make games more accessible. The downward trend in average prices indicates a clear orientation towards satisfying the demand for affordable and diverse games. Additionally, the persistent popularity of single-player games emphasizes the importance of narrative and individual exploration in user preference. Compatibility with various operating systems, especially Windows, Mac, and Linux, is identified as a crucial factor in expanding the potential user base, underscoring the importance of versatility in this aspect. Furthermore, Indie and Casual genres lead in demand, indicating a preference for accessible and easy-to-learn games.

To capitalize on these trends, it is suggested to develop games that are accessible to a broad audience, considering free or competitively priced options, multiplatform compatibility, and experiences for both single-player and multiplayer. Exploring emerging genres like Indie and Casual can offer opportunities, while adapting to popular languages such as English, simplified Chinese, and Spanish can significantly expand the audience. Leveraging the steady growth of the gaming industry is essential, focusing on creating high-quality games that meet the changing expectations of players.

The creation of visualizations in Power BI enabled a clear and effective presentation of the results, facilitating informed decision-making for those seeking success on the Steam platform.

The combination of data analysis, effective cleaning, and impactful visualizations has allowed us to contribute significantly to the understanding of this constantly evolving industry. While this analysis provides a solid foundation based on the analyzed data, there are still many variables to explore that may be significant for strategic decision-making by game developers, publishers, and companies seeking success in this dynamic and growing industry.