Name: Soo Min Hao

Student ID:20257525

Summary:

Steam, as the premier platform for digital distribution of video games, offers a unique and comprehensive ecosystem for understanding the dynamics of game success and user engagement.

Goal:

The project's primary goal using the Steam Games Dataset for creating data visualization and analysis techniques to uncover insights related to game pricing strategies, release trends, user review patterns, genre popularity, and the impact of development models like Early Access on game reception.

Objective:

1. To investigate the correlation between game pricing, both original and discounted, and the nature of user reviews.
2. To trace the evolution of game releases over the years and discern any patterns in pricing over time.
3. To explore how various game features, such as single-player and multiplayer options, affect user engagement and review volume.
4. To compare the performance of indie developers against major developers in terms of game pricing and user reviews.
5. To assess whether Early Access status influences the success of games as reflected by user reviews.

Target Audience:

- Gamers and gaming enthusiasts seeking a data-driven perspective on game trends and value.
- Game developers and publishers formulating or refining strategies for game release and pricing.
- Market analysts and researchers examining the digital distribution market for video games.

Problem Statement:
Several factors contribute to the popularity and reception of games on the Steam platform. What factors and elements like pricing, features, developer type, and release models intersect to influence a game's reception and longevity on Steam?

Dataset Details:

This dataset titled "**Steam Games Dataset.xlsx**" is obtained on the 27th of October 2023 and the latest update to the dataset was on the 20th October 2023 as of the time of obtaining the dataset. This dataset encompasses the following format and its attributes in sequence:

Attributes	**Description of Attributes**
Title	The name of the game.
Original Price	Initial listing price of game upon release.
Discounted Price	Price after discounts.
Release Date	Game's release date on Steam.
Link	URL to the game's Steam webpage online.
Game Description	Description provided by developers or publishers, often summary of the game
Recent Reviews Summary	Categorical summary of user reviews in recent times.
All Reviews Summary	Categorical summary of all user reviews since release date.
Recent Reviews Number	Number of reviews in the past 30 days.
All Reviews Number	Total user reviews since release.
Developer	Game's developer name // Company creating the game
Publisher	Company// Entity publishing and distributing the game.
Supported Languages	Languages the game is available in.
Popular Tags	Keywords indicating genre or features of the game.
Game Features	Specific features available in the game, such as multiplayer capability or VR support.
Minimum Requirements	The operating system the game supports, and its minimum system requirements needed to run the game.

Initial Questions with Description

Given the dynamic nature and the ever evolving and revolutionising gaming industry, my primary aim is to discern patterns and gain insights that can guide stakeholders in making informed decisions. The initial set of questions I aim to address with this dataset are as follows:

RQ1: **Games Released per Year and Pricing Evolution**

Are more games released every year compared to the previous years and how game prices evolved over time?

Description: This research question aimed to understand the growth pattern of the distribution of video game on the Steam platform. Specifically, it investigated whether the volume of games released has increased over the years, which could indicate market expansion or platform popularity.

RQ2: Price and User Reviews

How do user reviews correlate with game prices, both original and discounted and that do higher pricing lead to more critical reviews?

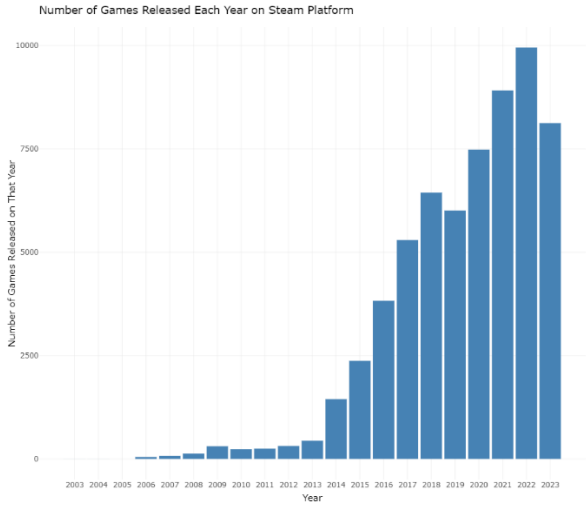
Description: This research question explored the relationship between game pricing and user reception on Steam. Do higher game prices correlate with more critical user reviews, possibly due to higher consumer expectations associated with cost? Conversely, it also considered whether lower-priced or discounted games received more favourable reviews due to perceived value for money.

RQ3: Developer and Publisher Analysis

Who are top 10 game developers or publishers most titles on Steam based on the total number of user reviews and what games' primary genre do they focus on?

Description: This question focused on identifying the top developers or publishers on Steam, based on the volume of user reviews received. It assumes that that a higher number of reviews might correlate with popularity or game engagement levels. Furthermore, the question explores the primary genres these successful developers and publishers specialize in, which could highlight genre trends and preferences within the gaming community on Steam.

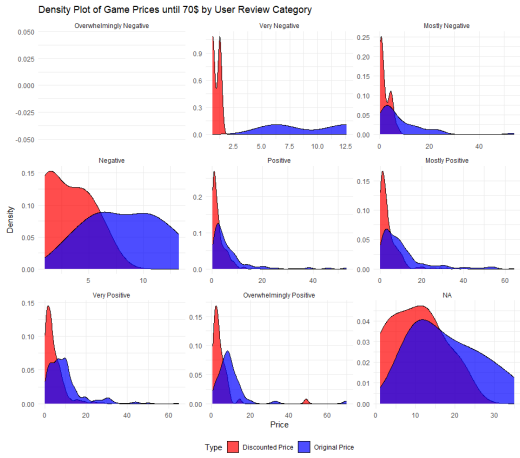
Dataset Assessment of Initial Questions



For RQ1, "Games Released per Year and Pricing Evolution," the above plot is a bar chart was used to illustrate the annual number of game releases on Steam, showcasing a clear upward trend from 2003 onwards. The bar chart was strategically chosen for its direct visual impact, where the height of each bar corresponded to the annual game count, making the rising trend a clear, uninterrupted view of the platform's expansion over two decades.

The dataset was cleansed to correct the release dates, and a decision was made to only include games released from 2003 onwards. This decision was made because Steam, the digital distribution platform, was launched in that year. While developers can add their games to Steam regardless of their original release date, it is filtered to avoid the distortion that could come from retroactive additions of older games, ensuring the analysis reflected the true year-to-year evolution of game releases and pricing on Steam.

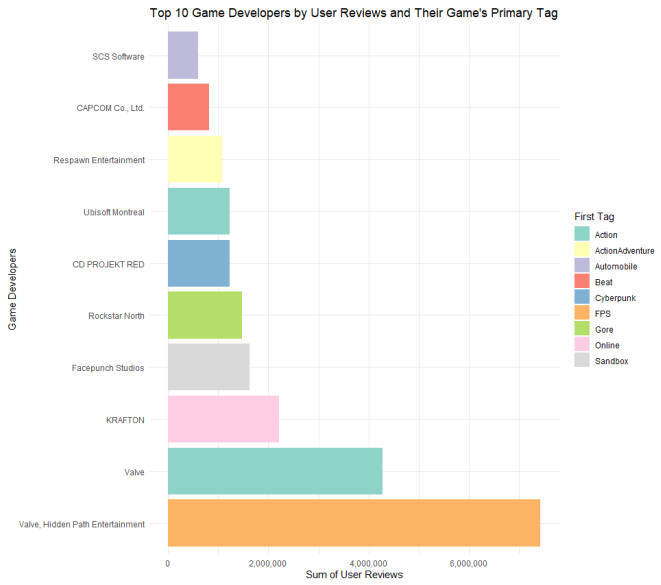
The above bar chart shows that more games are published on Steam yearly and with the exploratory visualization highlighted the platform's growth, confirms the hypothesis that there is indeed, growth in games released on Steam platform.



For RQ2, "Price and User Reviews," the investigation delved into the relationship between game prices and the reception by Steam's user community.

Using density plots, the visualization method effectively presents the distribution of game prices up to $70, categorizing them by user review sentiments. The reason until 70$ is used is because of the industry standard of game prices for AAA games. By focusing on a specific price range, the study avoided potential distortion from high-priced outliers, thus maintaining a more accurate visual representation. The analysis requires by thorough data cleaning, as for example: replacing prices from the format of $xx.xx(string) into numerical values.

The plot highlights a notable trend where discounted games tend to achieve more favourable reviews, suggesting that price reductions may positively influence user perceptions of value. Conversely, the density plots suggest that higher-priced games do not necessarily correlate with more negative reviews, challenging the notion that cost is a primary driver of critical feedback. This analysis underscores the nuanced interplay between pricing strategies and user satisfaction, inviting further exploration into how pricing impacts consumer behaviour and game success on digital platforms. Overall, this exploratory visualization contributes to a deeper understanding of the complex dynamics between game pricing and user reviews on Steam.



For RQ3, "Developer and Publisher Analysis," the objective was to identify the top developers on the Steam platform based on the aggregated number of user reviews and to identify the primary genres associated with their games. Performing this visualization requires one assumption, that all players are reviewers of their game and that first tag from ‘Popular Tags’ column is the most popular tag. Afterwards, the visualization approach was planned carefully in steps, starting with data cleaning by ensuring that developer names and numerical representations of user reviews are done correctly paired with an extraction of the most frequent game genre from the 'Popular Tags'—labelled as 'First Tag' for visual representation. The chosen visual encoding was a horizontal bar chart, which efficiently displays the hierarchical order of developers by reviews while simultaneously segmenting each bar by genre, using a diverse colour scheme for better and quicker genre recognition.

The horizontal orientation was especially advantageous for readability and space utilization, considering the length of developer names. This visualization provides a clear picture of which developers are leading on Steam. It is shown that both Valve and Hidden Path Entertainment are the number 1 developers based on their user reviews, but a game can have multiple developers. Hence if a game can have 1 developer, it's without the fact that Valve is the Number 1 Developer on Steam due to having both bar graphs named after.

Further Refined Questions with Description

RQ4: User Engagement in Singleplayer vs Multiplayer Games

Question: How do specific game features, such as having single-player and multi-player, influences user engagement and review counts?

Description: The inquiry centres on if different game features, specifically the availability of single-player and multiplayer modes, affect user engagement and the volume of review counts. This question touches upon the core aspects of gaming experiences and their social dimensions, which could influence a game's success and longevity in the market. By analysing user engagement metrics in relation to these game modes, this research could reveal player preferences and the potential impact of game features on player retention and community building.

RQ5: Indie vs. Major Developers:

Question: Given the increasing importance of indie games, how do indie developers' games perform in comparison to major developers in terms of pricing and reviews?

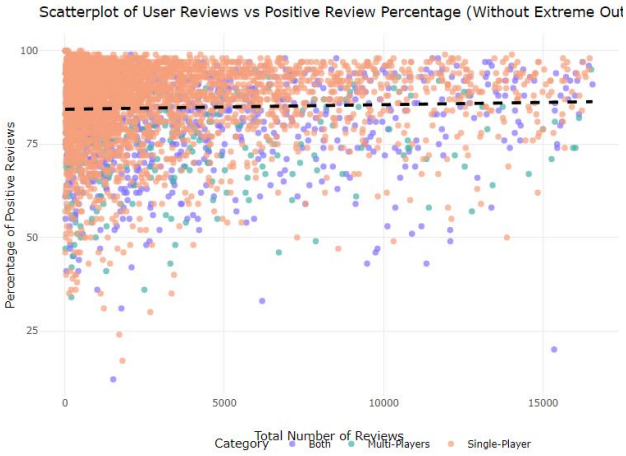
Description: This question explores the performance of indie developers' games as compared to those from major developers, with a focus on their games pricing and reviews. In an era where indie games are rising in popularity and presence, this study aims to shed light on whether indie developers can compete with major developers in terms of both the critical reception of their games and their pricing strategies. Insights derived from this could highlight the evolving landscape of the gaming industry and the role of indie games within it.

RQ6: Early Access Impact on Game Success

Question: Do games released in Early Access on Steam tend to have different user review outcomes compared to fully released games?

Description: Does Early Access model affect the eventual success of games on Steam. Early Access allows players to purchase and play a game while it is still in development, which can affect user perceptions and reviews. This question answers whether if an Early Access game leads to different review outcomes when compared to games that are fully released. Understanding this could be critical for developers considering Early Access as a pathway to full release, as it may influence marketing strategies, development timelines, and community engagement efforts.

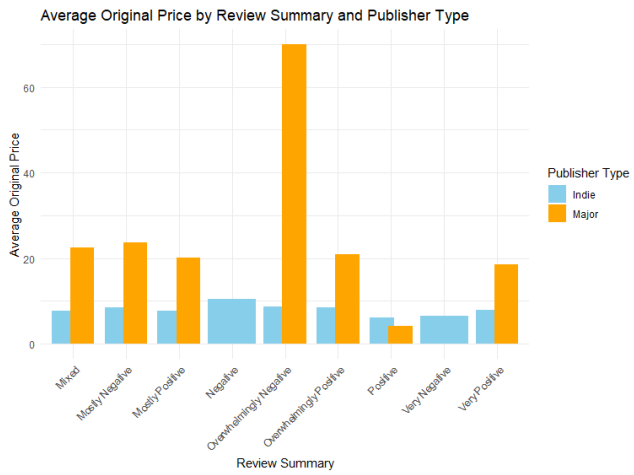
Dataset Assessment Further Refined Questions



For RQ4, "User Engagement in Singleplayer vs Multiplayer Games," this focuses on comparing how single-player and multiplayer features in games correlate with user engagement, as reflected by review counts and their positivity on Steam. This visualization strategy involved a data cleaning process, by combining any tags with ‘Online’ into Multiplayers and filtering out having games that have neither tag to ensure that the categories for game features were accurately captured and that converting reviews from strings to only numeric. The transformation included creating a categorical variable that distinctly marked games with single-player, multiplayer, or both features.

The visualization uses a scatterplot to juxtapose the number of reviews against the percentage of positive reviews, categorized by game type. This plot type was chosen for its ability to display a vast amount of data points, facilitating the observation of trends or outliers within specific categories. Using logarithmic transformation on review counts, the outliers can be mitigated, and the skewness caused by games with exceptionally high review counts can be removed for a normalized data. The color-coding by game type added a layer of differentiation, making it visually immediate to discern between the game types.

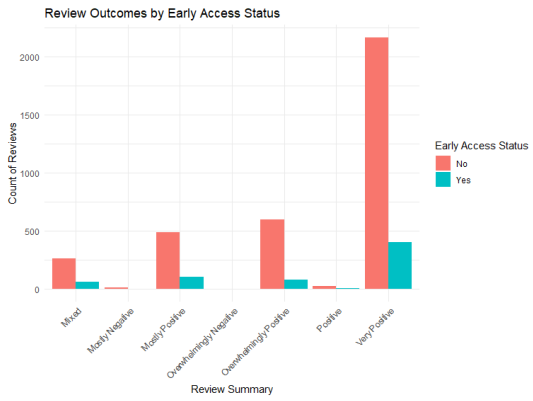
This visualisation proves that Singleplayers’ games and titles are more on lower review counts due to it not being a social and multiplayer games which requires more users for the games to be active, hence multiplayers games tend to have more user reviews. By excluding extreme outliers, the scatterplot provides a more representative and interpretable visualization of the typical range of user engagement across game types.



For RQ5, "Indie vs. Major Developers," the research aimed to compare the performance of indie and major developers through the lens of pricing and user reviews. The visualization starting with data cleaning, to differentiate between indie and major developers based on a predefined list of major publishers. Prices were normalized by removing the currency symbol and converting to numeric values, and reviews were categorized for comparative analysis.

The chosen visualization was a grouped bar chart to depict average game prices by user review categories for indie versus major publishers. This format was selected for a great contrasting view on the two publisher types across multiple review categories, making disparities immediately apparent. The colour scheme as well allows for quick identification of indie and major publishers.

The analysis revealed that indie games are competitively priced and often receive positive reviews, holding their ground against major publishers. The visualization highlighted that major developers' games occasionally have higher prices, particularly in categories with extreme reviews. This could imply a correlation between higher expectations and cost, or it could reflect the premium branding of major developers, which explains the huge graph in the negative receiving due to the higher cost, for a higher expectation.



For RQ6, "Early Access Impact on Game Success," the study focused on the influence of Steam's Early Access program on user review outcomes. The visualization process began with data cleaning to categorize games by Early Access status, followed by filtering review summaries to include relevant user feedback. By converting review counts from text to numeric, the analysis ensured a reliable quantitative basis for comparison.

The visualization uses a grouped bar chart, for a more contrasting review outcomes between Early Access and fully released games. The bar chart was an appropriate choice for its clear representation of the count of reviews across different summary categories, allowing for an immediate visual comparison of user feedback trends.

The analysis showed that Early Access games have a distinct review pattern compared to fully released titles, particularly in the 'Very Positive' category, suggesting that Early Access may foster a more engaged or lenient community. However, extreme review categories like 'Overwhelmingly Negative' showed a substantial count for non-Early Access games, indicating that fully released titles might face harsher criticism upon failure to meet expectations.

Reflection on the development process

Reflecting on this project, I am struck by the depth and complexity of the Steam dataset due to it not being clean and contains messy and ‘dirty’ data and the complexity nature of the analysis required to glean meaningful insights from such a dataset. The journey through the dataset has been both challenging and enlightening, marked by moments of stress, repetition, and frustration, but also by discovery and learning. The dataset used is not fully reflected on the exact numbers of games released on steam as games who are released before 2003 when steam wasn’t available could be added into the steam library by developers. The discounted prices of games in the dataset reflect the discounts available as of 27th October 2023 and therefore, Steam frequently has sale events, and the game prices obtained might be influenced by any ongoing or recent sales. The specific sale events around this date are not detailed in the dataset, therefore this context should be considered when interpreting any insights related to pricing.

The process of cleaning and transforming the data for each research question often felt tedious and full of works and frustration, as it involved high attention to detail and the handling of various exceptions and edge cases. Many repeated data checks and the iterative nature of refining visualization techniques sometimes led to frustration, especially when

encountering unexpected problems, such as inconsistent data entries or when standard visualization methods did not yield the clarity of insight I sought. Creating visualizations that accurately represent complex data trends requires high patience and creative problem-solving.

Each research question brought with it a unique set of challenges. For instance, distinguishing between indie and major developers required a careful definition of what constitutes each category and a consideration of how these categorizations would influence the analysis. Similarly, the exploration of Early Access games and their impact on user review outcomes necessitated a nuanced approach to account for the different stages of game development and user expectations. Throughout the project, certain tasks became repetitive, such as the necessity to convert strings of review counts and percentages into numeric data that could be graphically represented. Despite this, there was a certain satisfaction in seeing the data transform into actionable insights.

The stress induced by tight deadlines and the expectation to produce meaningful outcomes from the data was palpable. Yet, these pressures also fuelled a rigorous exploratory process, where new questions and visualization methods emerged organically from the data. This project requires many iterative natures of data analysis, where each step builds upon the last, and flexibility is key to adapting to new findings. In reflection, this project was as much about managing the patience of a programmer and data analyse as it was about the technical aspects of data science.