CSC 47400 Visualization

Fall 2025

Steam Explorer Final Project

Prof. Yunhua Zhao

Due: December 17th, 2025

Jezlea Ortega & Esther Mallen

Abstract:

Indie game developers are facing a major discoverability challenge in game purchasing platforms, specifically Steam. Thousands of new titles are released each year on Steam, decreasing visibility for many high-quality games. Due to Steam's unclear recommendation and tagging systems, developers are unable to understand the various factors, such as genre, tags, price, release timing, and user reception, that affect a game's visibility. Without a clear understanding of these factors, discoverability issues remain unsolved in the gaming industry and continue to negatively impact developers, especially small studios with limited marketing resources. This project aims to address the discoverability challenge on Steam plaguing small studios in particular by creating an interactive dashboard that explores multivariate relationships in a large Steam games dataset. Through the implementation of visual views, filtering, and details on demand, our dashboard will help users identify patterns related to increased and decreased discoverability. These patterns will include trends related to which game characteristics are associated with higher visibility or underserved niches. Overall, the goal of this project is to provide developers with a data-driven tool to better understand the marketplace and make informed decisions regarding how to market and position their games.

Problem Description:

The problem this project addresses is the discoverability crisis impacting indie game developers on Steam. In the past decade, Steam has experienced unprecedented growth in the quantity of games released. According to SteamDB, industry records note the number of annual releases rising from a few hundred in the early 2010s to more than 18,000 games in 2024, making recent yearly output exceed the platform's entire first decade (SteamDB). This over-saturation on the platform makes it significantly more difficult for smaller studios with

newly released games to achieve visibility in an increasingly crowded marketplace. Not to mention, prior research consistently highlights that visibility on Steam is not determined solely by game quality, but by a complex combination of factors including genre competition, tag selections, release timing, pricing, and user reviews. As stated in “The Impact of Experience The Influences of User and Online Review Ratings on the Performance of Video Games in the US Market” by Sven Joeckel, game visibility is influenced by a complex set of interacting factors like ratings and reviews: “games that sell well also are rated by a higher number of users and games that are rated by a higher number of users are also correlated with higher user scores while user scores and press scores are related with each other” (Jöckel 2011). This demonstrates, highly recommended and visible games have a direct correlation to also maintaining a high quantity of reviews and significantly greater user scores. Additionally, the academic article, “Empirical investigation of key business factors for digital game performance” by Saiqa Aleem, Luiz Fernando Capretz, and Faheem Ahmed, identifies the correlation between price, release timing, genre competition, and marketing resources in a games traction: “The path coefficients of five of the seven variables (customer satisfaction, market orientation, time to market, monetization strategy, and brand name strategy) were positive and found to be statistically significant at $p<0.05$” (Aleem, Capretz, and Ahmed 2015). This shows, customer satisfaction, market orientation, time to market, monetization strategy, and brand name strategy play a major role in the success of a video game.

Moreover, due to the complex combination of factors correlating to high visibility games and Steam's lack of transparency in their internal discovery algorithms, developers face a major dilemma in determining how to structure their game to achieve maximum success. Particularly, the struggle is emphasized for indie developers due to various factors, including

their limited marketing budgets, lack of publisher support, lack of pre-existing fanbases, and high competition with other genres. The lack of clear insight into how attributes influence exposure prevents indie developers from making crucial and costly decisions with data-driven reasoning. According to “Shape patterns in popularity series of video games”, this struggle is exacerbated through games following a popularity trajectory of brief visibility before a rapid decline, “Cluster 1 (blue), which accounts for 47.4% of the games, displays a predominant pattern of decay, making it clear that a significant portion of games reaches its peak popularity in the first few days after launch and then gradually fades over time” (Cunha, Pessa, and Mendes 2024). In turn, this amplifies the importance for indie developers to understand the early conditions that increase or limit discoverability. Although existing resources such as SteamSpy or SteamCharts attempt to rectify this issue through partial insights into game metrics, they do not provide an interactive, multivariate analysis of the relationships between game attributes and discoverability. The lack of this type of analysis, in addition to the inability to visualize how combinations of factors, like tags, price, user sentiment, and genre density, interact at scale, prevents developers from determining visibility-increasing or limiting patterns.

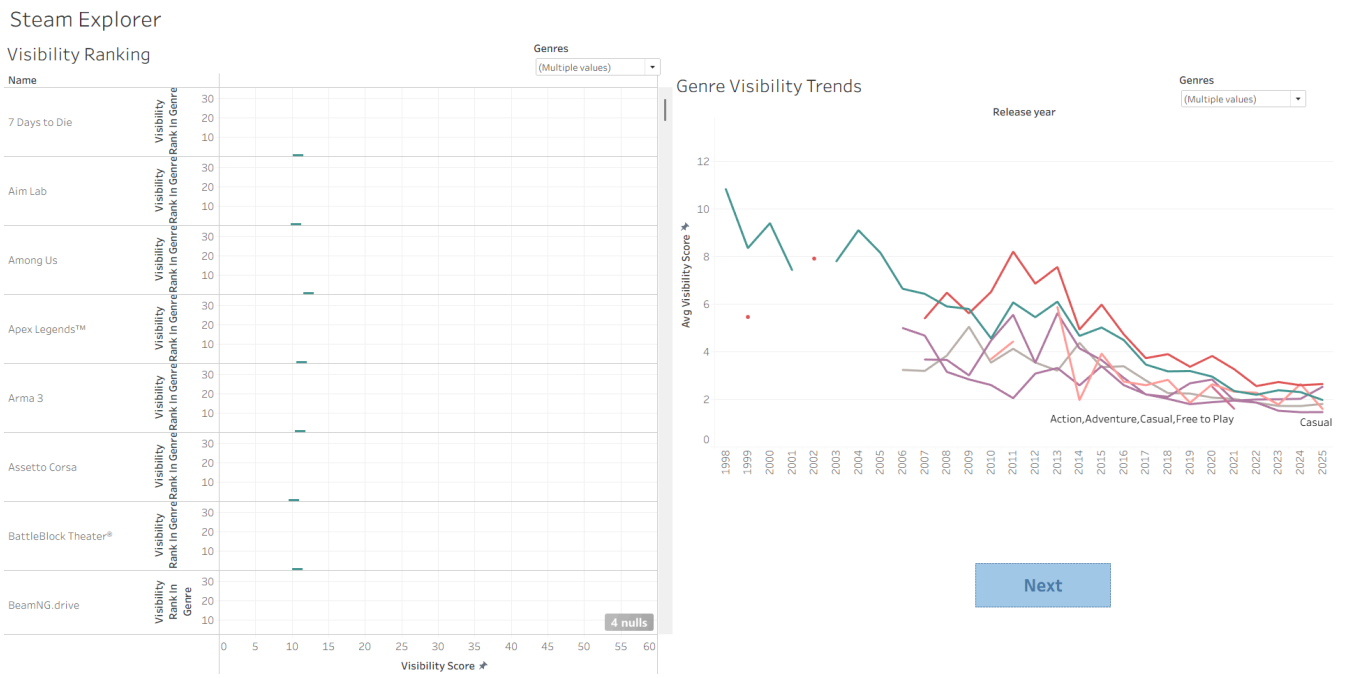
Data Processing Python Code:

This project utilized the Steam Games Dataset on Kaggle by Martin Bustos, a large real-world collection of metadata about video games available on the Steam platform. Furthermore, Jupyter Notebook and Python were implemented to preprocess the data, analyze it, and create tables for Tableau. Data manipulation and intermediate output saving in CSV files were implemented by also using Pandas and numpy. The code encompassed four functions with an additional main() function that calls each: load_and_clean, engineer_features, build_aggregates, and discoverability. The load_and_clean function reads the raw CSV file,

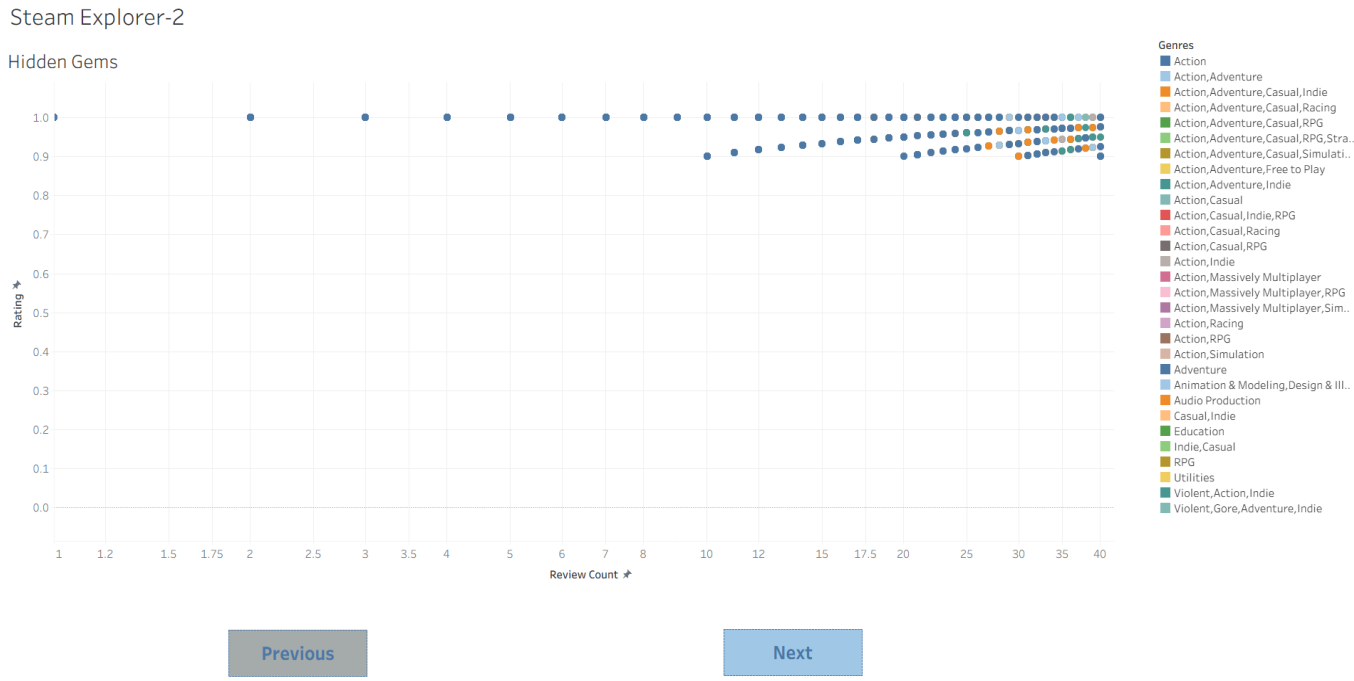
keeps only the relevant columns, fixes misaligned fields, converts numeric strings to numeric types, and extracts the release year from the date. The rows with missing game titles are also dropped here and missing numeric values are replaced with safe defaults, before writing the cleaned dataset to processed/steam_clean.csv. Additionally, the engineer_features uses the positive and negative review counts to compute review_count (total reviews) and rating (positive share). Then, log_review_count and the visibility_score = rating * log(1 + review_count) is calculated to approximate the particular game's visibility score on Steam. Similarly, the function determines tag_count (number of tags a game has), is_indie flag (if a game has the indie genre), the price_bucket (very cheap, cheap, mid, expensive), and age years (how old a game is based on the difference between the current year and release year. The results are then saved to processed/steam_features.csv. Afterwards, build_aggregates groups the games by genres to compute the number of games, average rating, and average visibility score per genre. The function also aggregates by release year to get yearly counts and mean ratings, and by (Genres, Release year) pairs to analyze genre composition and visibility. The results are written into separate CSV files. Lastly, the discoverability function uses the feature-engineered dataset to reveal any hidden gems. The function computes the visibility_rank_in_genre by ranking games within each genre based on their visibility score. Then, it filters for titles with high ratings, but low review counts (for example, rating ≥ 0.9 and ≤ 40 reviews). The results are saved in discoverability/hidden_gems.csv and the top 200 most visible games overall in discoverability/top_visible_games.csv.

Tableau Interface Design:

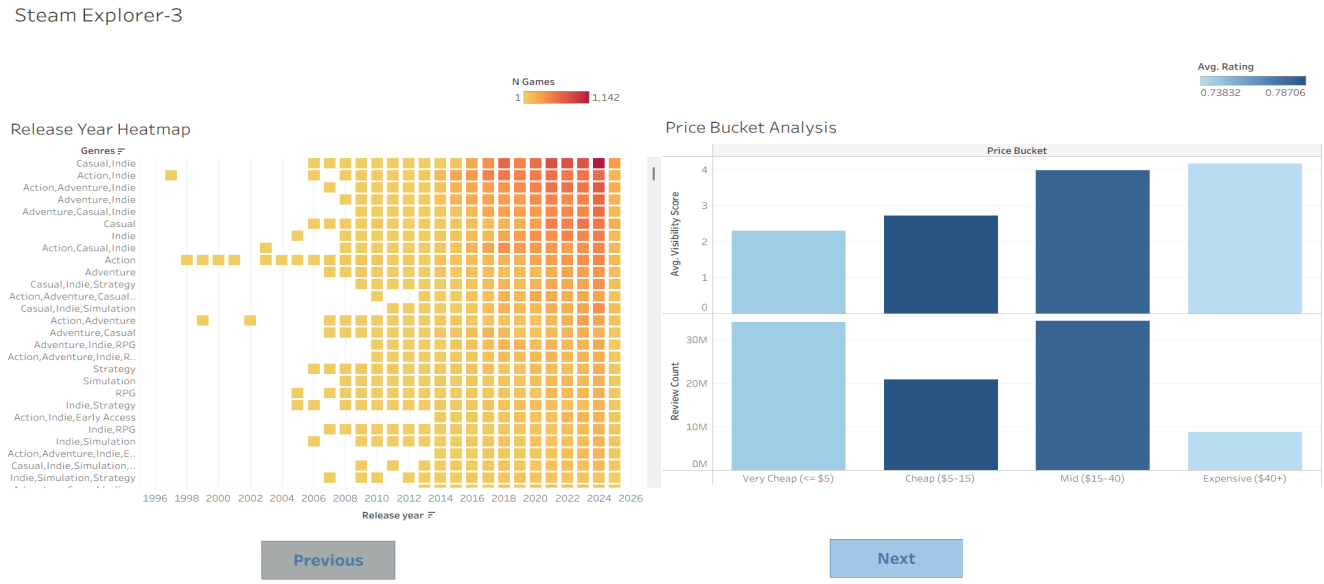
Page 1: Visibility Ranking for Games and Genre Visibility Trends in Genres



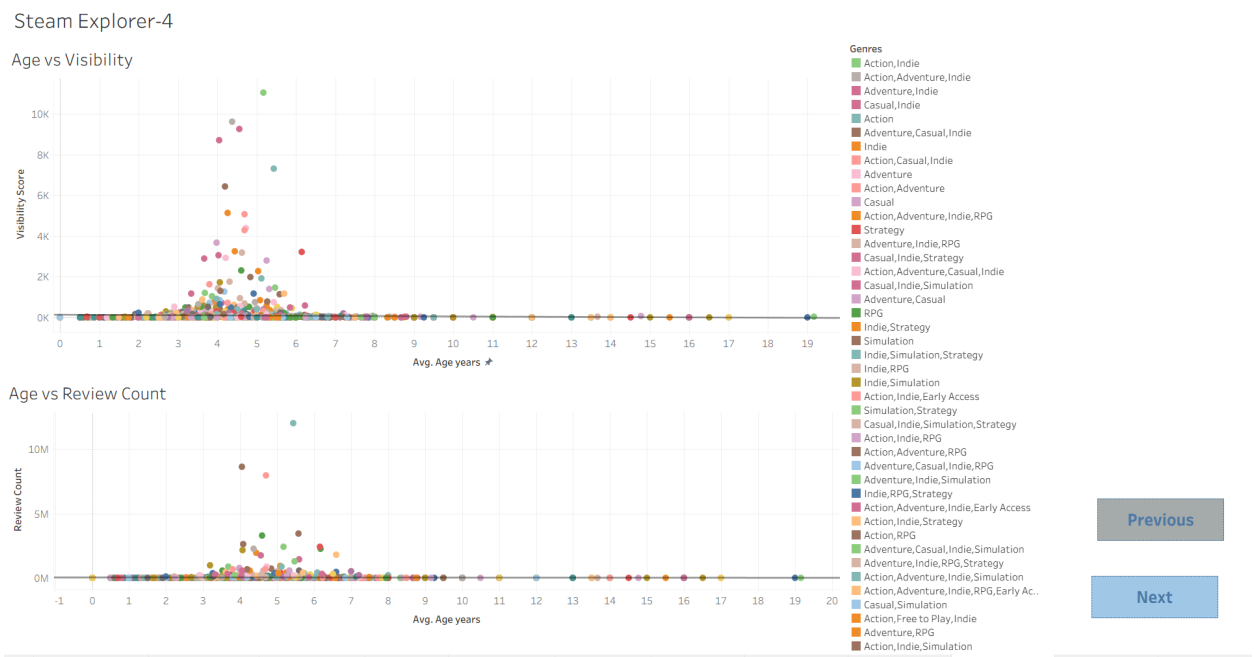
Page 2: Hidden Gems



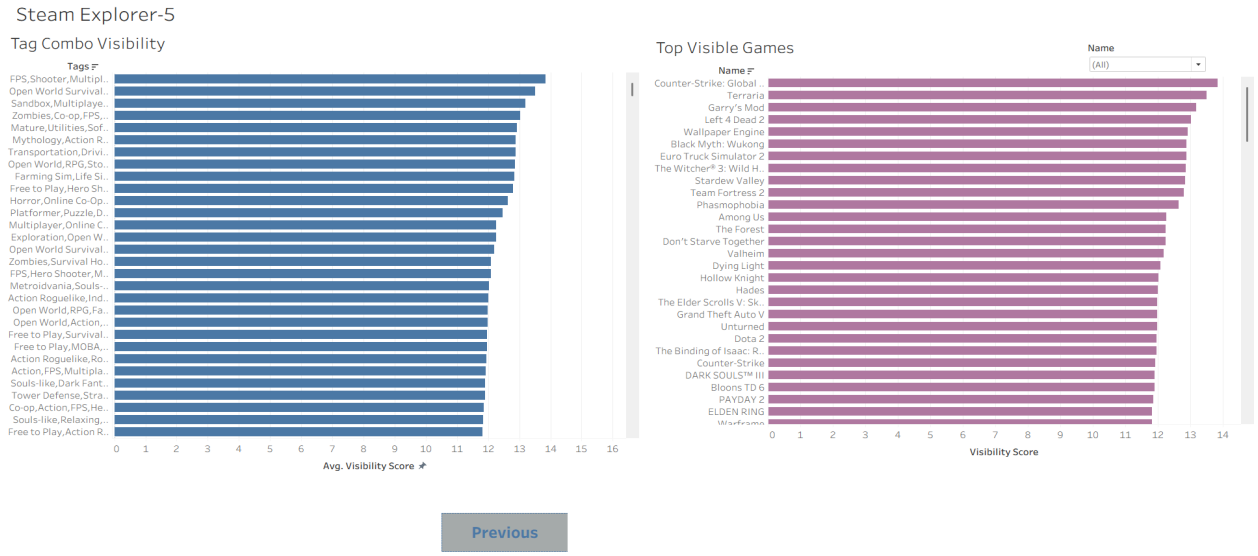
Page 3: Release Year Heatmap and Price Bucket Analysis of Genres



Page 4: Visibility and Review Count against Age for Genres



Page 5: Tag Combinations for Visibility and Top Visible Games



Findings / Use Cases:

After plotting the data and analysing the charts, we found that Steam prioritises review count and ratings for game visibility, even for older games, but they still need to have a lot of activity. If people are not playing a game enough, even if the ratings are good, they get buried on the platform, not getting any visibility. Also, over the years, the visibility of games based on genres decreases due to oversaturation. Casual/Indie is the genre with the highest number of games, and it falls into the Cheap price bucket, which is 5 to 15 dollars. This is important for future creators who want to know the factors that increase visibility for their game. We can conclude that Steam's system can favor new games as long as they receive a lot of reviews and activity. This interactive dashboard can be a great resource for users to find good games with low visibility and for developers to take in consideration what can make their games more visible to the users.

Limitations & Future Work

Some limitations we encountered while doing this project were that some data got lost while creating the visualizations. Tableau had some problems with some data points, and they were showing up as null even though they had data in the columns. It also has a data size limit of 1000 for some charts, or even less. In terms of the filtering for the charts, some don't work depending on the type of charts.

Things we considered that can be implemented or fixed in the future include creating a UI/UX interface for the visualizations. Due to confusion in class, we were not aware that we needed an actual website for the interactive dashboard, so we did an interactive dashboard on Tableau instead. If we create a website, we can deploy the site, and people can interact with our visualizations. Also, we were thinking about using a different tool for the visualizations since Tableau has some limitations that decreased the performance of our charts.

Bibliography:

- Aleem, S., Capretz, L. F., & Ahmed, F. (2015, November 13). Empirical investigation of key business factors for Digital Game Performance. arXiv.org. https://arxiv.org/abs/1511.04422
- Cunha, L. R., Pessa, A. A. B., & Mendes, R. S. (2024, June 4). *Shape patterns in popularity series of video games*. arXiv.org. https://arxiv.org/abs/2406.10241
- Steam database · steamdb. (n.d.). https://steamdb.info/
- (PDF) the impact of experience the influences of user and online review ratings on the performance of video games in the US market. (n.d.-a). https://www.researchgate.net/publication/228642165_The_Impact_of_Experience_The_Influences_of_User_and_Online_Review_Ratings_on_the_Performance_of_Video_Games_in_the_US_Market
- Bustos , Martin . “Steam Games Dataset.” *Www.kaggle.com*, (2025, April 25), www.kaggle.com/datasets/fronkongames/steam-games-dataset?resource=download.