**SUMMARY**

Steam is a global application with massive user base that serves as the biggest platform for digital distribution of video games.

**Goal:**The goal of this project is to analyse and visualize the Steam Games Dataset to form insights about game pricing, popularity, user reviews, and other metadata, ultimately aiming to understand the trends and factors influencing the success or failure of a game on Steam.

**Objective:**

1. Examine the relationship between game pricing (both original and discounted) and user reviews.

2. Identify trends based on the release date of games.

3. Understand the correlation between game features, popular tags, and game success.

**Target Audience:**

Gamers and Gaming Enthusiasts: To discover trends and patterns in the gaming world.Game Developers and Publishers: To understand what drives success and user interaction.  
Market Analysts and Researchers: To study the digital game distribution landscape.

**Problem Statement:**

Despite the vast number of games available on Steam, not all achieve the same level of success or wider users reach. What factors influence a game's popularity and user reception on Steam? How do pricing strategies, release dates, game features, and other attributes play into this?

**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 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:

|  |  |  |
| --- | --- | --- |
| No | Initial Set of Question | Initial Questions Description |
| **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? |
| **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? |
| **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? What games’ primary genre do they focus on? |

**Description of Visualization Strategies Used**

|  |  |  |
| --- | --- | --- |
| Strategies | **Description of Visualization Strategies Used** | **Applicable RQs** |
| **Data Loading:** | -Used ‘readxl’ library to load the steam game dataset from the Excel file. | All RQs |
| **Data Cleaning** | -Removed the "UTC" format from the "Release Date" column.  -Convert the column to the date format using the lubridate and gsub functions. | RQ1 |
|  | -Handled missing 'All Reviews Summary’.  -Converted 'FREE' in 'Original Price' and 'Discounted Price' to numerical zero. | RQ2 |
|  | -Removed company suffixes and unnecessary symbols from developer names to ensure consistency | RQ3 |
| **Data Transformation** | -Filtered out games released before 2003. | RQ1 |
|  | -Extracted 'GameDevelopers' from 'Developer' column after cleaning it.  -Converted 'All Reviews Number' into numeric values for aggregation. | RQ3 |
| **Data Aggregation** | - Grouped by year of release; joined with reference frame for complete year data. | RQ1 |
|  | -Grouped reviews by category to understand the distribution of prices. | RQ2 |
|  | -Aggregated user reviews per developer and computed the sum to identify the top 10 developers. | RQ3 |
| **Visual Encoding** | -Bar chart: x-axis: years  y-axis: number of games released on that year. | RQ1 |
|  | -Box plots:  1. Original Price vs All Reviews Summary  2. Discounted Price vs All Review Summary | RQ2 |
|  | -Mosaic plot used to represent the concentration of user reviews among top developers along with their most popular game tags. | RQ3 |
| **Theme & Aesthetics** | -Bars represented in a "steelblue" colour for number of games released. | RQ1 |
|  | -Highlighted outliers within the boxplot with black dots. | RQ2 |
|  | -Applied a colorblind-friendly palette with a minimal theme for clarity and readability. Rotated x-axis labels and flipped coordinates for better label visibility. | RQ3 |
| **Interactivity** | Converted ggplot chart to interactive using plotly. | All RQs |

**Exploratory Process  
RQ1  
Initial Inspection:**

Upon initially inspection of the dataset, it was observed that the "Release Date" column had "UTC" appended to the dates, which would prevent accurate date conversions and calculations.

**Defining the Scope:**

Given that steam was released in 2003 and the dataset might contain games released before the year 2003 and developers can add games into Steam library regardless of release date, I decided to narrow down the scope and focus on games released on Steam after 2003.

**Visualization Choices:**

I chose a bar chart as it’s effective in representing the frequency of the number of games released per year as bar chart is clear in visualising trend over the years and allows for easy comparison between years.

**Iterative Process:**Initially I wanted to add the average price of each game and comparing the prices however this question is straighter to point and hence I chose number of games released per year. After the initial visualization was created, I made it more interactive for easier usability.

**RQ2**

**Initial Inspection:**

Upon initially inspecting the dataset, I noticed there were multiple columns related to pricing and user reviews. The presence of both original and discounted prices, as well as an array of review categorizations, opened the avenue for an in-depth exploration of the correlation between game pricing and user feedback.

**Defining the Scope:**

Recognizing that game prices vary widely based on multiple factors like popularity, game developer reputation, game type, and more, I decided to understand if there was any discernible relationship between how the game was priced and the user reviews it received, especially in cases of discounts.

**Visualization Choices:**

I decided on a combination of box plots and scatter plots. Box plots are effective in presenting distributions of prices across different review categories, allowing for easy identification of median values, spread, and outliers. Scatter plots added granularity by plotting individual game prices versus their review categorizations.

**Iterative Process:**

Initially, my aim was to purely understand the distribution of prices. However, as I delved deeper, the layer of user reviews added another dimension. It allowed me to understand not just how games were priced, but how those prices might influence user perception and feedback.

**Initial Inspection:**

When inspecting the dataset, it became clear that developers varied widely in terms of the number of user reviews they had received. It was necessary to convert textual review counts into numerical data to facilitate further analysis.

**Defining the Scope:**

The objective was to ascertain the most prominent game developers or publishers on Steam, focusing on the number of user reviews they amassed, their primary genres, and the general reception of their games.

**Visualization Choices:**

A mosaic plot was chosen for its ability to showcase the distribution of user reviews across different developers while also indicating the most popular genre associated with each developer.

**Iterative Process:**

The initial step was straightforward: identifying the top developers based on user review counts. Then, the exploration involved deciphering the most popular tags associated with these developers, which involved splitting, counting, and grouping tags. The final visualization was refined for clarity by adjusting themes and aesthetics.

(Before critical discussion) \*\*5. Further refine/propose questions\*\*:   
  
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?

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?  
  
-\*\* RQ8\*\*: System Requirements and Audience Reach: How do games' minimum system requirements relate to their popularity and user reviews? Is there a correlation suggesting that games with lower system requirements garner a wider audience or better reviews?

**Critical discussion of visualization design**

**Reflection on the development process**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, it’s worth noting that 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.