

# Technical report of the project on Introduction to Data Science

Developers: Linsen Gao, Sergi Panarin, Max Väistö

Project Title: [SteamSavvy](#)

## 1. Motivation

The primary objective of our project, SteamSavvy, is to harness data from the Steam gaming platform to deliver valuable insights to a range of users, including investors, developers, marketers, and advertisers. These insights provide a comprehensive view of the gaming industry, allowing stakeholders to make informed decisions about market trends and the overall performance of companies within the industry.

Our project is structured around two core elements: game genre comparison and game company performance metrics. In the game genre comparison section, we aim to provide clarity within the gaming world by analyzing which genres are thriving, which are on the decline, and which are maintaining their popularity. These insights empower developers to create games that meet current demand and aid investors in making well-informed decisions. In the company performance metrics section, we assess how various game developers and publishers are performing on Steam, offering insights into both established industry leaders and emerging talents.

## 2. Technical processes

Our technical approach consists of several key steps, including data gathering and wrangling, exploratory data analysis (EDA) and visualization, a machine learning approach, and the development of a user interface (UI) platform.

### 2.1 Data gathering and wrangling

We collected game data from [SteamSpy](#) and supplemented it with data from the Steam API. From SteamSpy, we obtained information on over 60,000 games. This data included categories such as publisher, developer, score rank, positive and negative review counts, owner numbers, current price, initial price, discounts, concurrent user counts, available platforms, release dates, game categories, downloadable content (DLC) details, available languages, genres, tags, and review ratings. Additionally, we gathered detailed information about game DLCs in a separate CSV file, which included details on the DLC developer, publisher, review counts, owner numbers, price, and release date. However, we did not utilize the DLC data in our subsequent analysis.

To prepare the data for analysis, we conducted data wrangling and cleaning, which involved removing irrelevant, null, or duplicated data. We parsed the data from JSON format into a data frame, ensuring consistent data formatting for further analysis.

### 2.3 EDA and Visualization

Exploratory Data Analysis (EDA) is a crucial step in our project to understand the characteristics of the data and to draw meaningful insights. We presented the distribution of data from various categories. Plotly was used to create interactive figures for the end product, which features a dashboard for user interaction.

In the genre performance section, we employed bar charts to visually demonstrate the relative differences between various game genres. This was achieved by estimating genre popularity and the number of games developed in the next two years. Bar charts were used to illustrate the popularity and revenue share of different genres.



Figure 1. Genre performance visualization

We also conducted an analysis of user ratings for both free and non-free games based on the number of game owners. The distribution curves and rug plots were presented in a visualization figure.

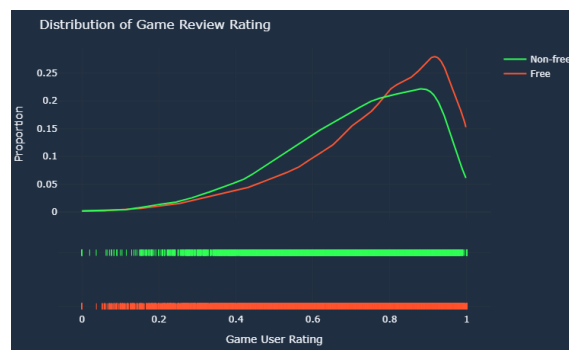


Figure 2. Free vs non-free review distribution visualization

Furthermore, we explored the relationship between the number of game owners, revenue, and the number of games released by different companies. This relationship was depicted using bubble plots, which provided insights into the market performance of each company.

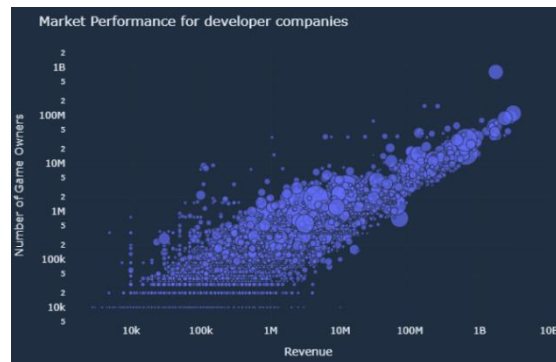


Figure 3. Free vs non-free review distribution visualization

We also incorporate a visualization component that presents comprehensive data on the company you choose by developer or publisher, encompassing game sale revenue estimates, released game numbers, concurrent user numbers, popular game genres, and the average game ratings. A bar chart is included to illustrate the annual number of games released by the company.

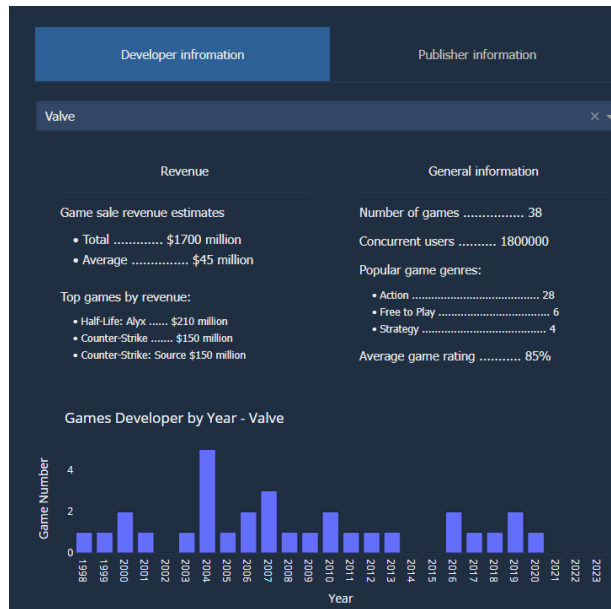


Figure 4. Company information visualization

All the data points within the figures offer interactive functionality through our website. For instance, in Figure 1, you have the option to selectively exclude a legend in the bar chart. In Figure 2, you can apply filters to adjust the number of game owners and specify a minimum threshold for reviews. Additionally, Figure 3 allows you to modify the displayed game count.

## 2.4 Machine Learning

Before the algorithms for Machine Learning could be applied, data needed rearranging and cleaning. A small number of entries had empty or null values in certain columns. Due to the nature of the data, it was decided to ignore these few rows from the database, as it was impossible to reliably estimate their owner counts and release dates. Luckily, none of them were top selling games, and thus they could have been omitted for the sake of this analysis. Release dates were also transformed into ordinal values where necessary for computation purposes.

After data had been prepared, it was then loaded into pandas dataframes, separated by genres and stored as python dictionaries, where genres are represented as keys and values are dataframes containing all relevant information about games of that genre. Since many games have multiple genres, they can be used several times. After processing, the data was ready to be analyzed.

We created a time-series, where we analyze the number of games' owners and numbers of games released in 2-month intervals. This data is then fed to the Regression model, which we use to predict what are the likely values for the next two years. For the moment, the project uses Linear Regression to give the overall trend of the industry. Regression lines also help us to analyze possible opportunities, where we use the slope ratios to see which genre has the best ratio of number of players vs number of games. If the genre has many players but a low number of games, it presents an opportunity for the market.

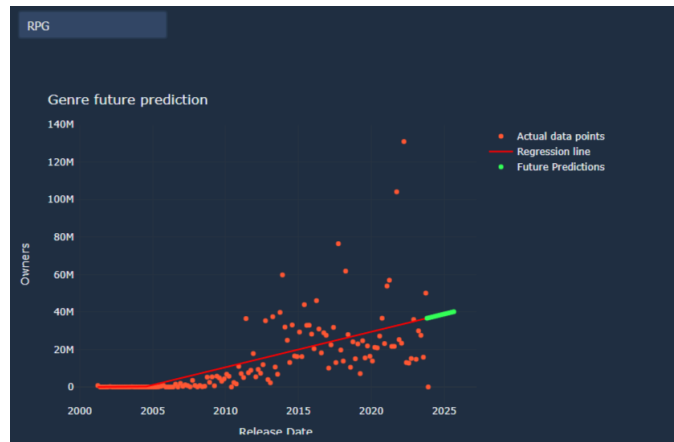


Figure 5. RPG genre regression analysis

## 2.5 Building the platform (UI)

During the entire duration of our project, we were working on creating an interactive website, leveraging the capabilities of Plotly and Dash. The dashboard, a core feature, demanded significant time and effort. This was our first look into website development, which involved hand-crafting the User Interface. In hindsight, crafting the UI by hand might not have been the most efficient approach if there were superior alternatives.

The beginning of UI development was simultaneous with data accumulation. However, as we integrated more features, we stumbled upon performance bottlenecks. The server often timed out due to prolonged data loading and processing, resulting in a non-responsive web session.

We had to change our approach to data processing. Instead of on-the-fly processing during user interactions, we pre-processed data before initializing the application, utilizing the pickle library for storing intermediary files. This alteration boosted the system's performance by over six times. Further upgrades were achieved by transitioning from standard Python objects like lists to the more efficient NumPy arrays, doubling the speed once again. Key insights from ChatGPT significantly aided this optimization, particularly when given directions on library preferences for function reimplementations.

While the Plotly and Dash frameworks inherently impacted the system's speed, we've incorporated loading icons as indicators, ensuring users are aware of ongoing data retrievals. Overall, optimizing the project demanded extra effort, but the final product now offers a smoother and more responsive experience to users.

## 3. Summary

The project's value lies in providing enhanced insights into Steam platform trends, benefiting game company higher management, investors, advertisers, and so on. It enables these stakeholders to gain valuable information about market trends and obtain a comprehensive overview of company performance, ultimately empowering them to make data-driven decisions in the dynamic gaming industry.

In the original project plan, our scope encompassed the analysis of additional categories, including release dates, platforms, CCU data, and DLC information. However, as we progressed through the EDA phase and developed a deeper and clearer understanding of our objectives, we decided not to pursue these specific aspects.

One of the ways to further improve our project is to introduce different non-linear models to our ML analysis. It is significantly more challenging, as picking the correct model requires a great deal of insight, and creating an automatic tool for determining the right model is not trivial either. If done correctly, this feature can be monetized

to give the best analysis possible to the most loyal customers. It can also be used as a fun tool to see which prediction models yield interesting results.