# Abstract

The video game business is a large industry that produces, promotes, and commercialises video games. It is also one of the world's foremost entertainment industries, generating more income than the film and music industries combined. The main objective is to use predictive analytics to predict the success of upcoming video games based on user engagement data with maximum accuracy. A model is being proposed to explore and analyse the user data, which likely consists of variables such as reviews, ratings, revenue and other attributes covering all data types such as Nominal, Ordinal, Ratio and Interval. The framework will include clustering, classification, and regression analysis methods through data collection and pre-processing. This model will use the CRISP-DM framework to forecast the sales of upcoming PC games using historical data. The final model will be deployed to estimate game sales as an outcome, with iterative updates to improve accuracy.

Table of Contents

[Abstract 1](#_Toc184075608)

[Table of Contents 2](#_Toc184075608)

[Introduction 3](#_Toc184075609)

[Data gathering and data analysis 5](#_Toc184075613)

[A. Feature Selection 5](#_Toc184075614)

[B. Data Gathering 5](#_Toc184075615)

[C. Data Analysis 7](#_Toc184075616)

[Conclusion 9](#_Toc184075620)

[Bibliography 10](#_Toc184075621)

[Appendices 11](#_Toc184075622)

# Introduction

# Business Problem Statement

Video games provide interactive entertainment where players interact with audio-visual elements displayed on the screen using input devices such as controllers, keyboards and mice (“Video Game,” 2024). The video game industry is a billion-dollar generating business and has been there for many years. In 2024, the revenue from the global gaming market was estimated to be 455 billion USD; the mobile gaming market generated an estimated 98.7 billion USD. Despite the revenue, the gaming industry faced major misdirection (J. Clement, 2024b). After COVID-19, the market decline was more substantial than expected for AAA games, with economic and political factors strongly impacting the users.

Recently, Video games have become expensive to make both in terms of development and distribution, making economic consumers buy fewer games at total price (around 70£ per game). Furthermore, many publishers such as Epic Games, HoYoverse, and Bungie have engaged their audiences with Games-as-a-Service (GAAS) titles such as Fortnite, Genshin Impact, or Destiny 2. With innumerable hours invested in a specific GAAS title that receives a steady stream of updates frequently, gamers are less likely to switch to new titles (J. Clement, 2024b). Most GAAS titles are Free-to-Play, which means there is no entry fee to play these games, and they make the most revenue through [microtransactions](https://www.statista.com/statistics/274761/electronic-arts-ea-extra-content-revenues/) compared to AAA games, which require an entry fee. For instance, Electronic Arts (EA) generated substantial money in 2024 from live service games, earning over USD 4.46 billion from [microtransactions](https://www.statista.com/statistics/274761/electronic-arts-ea-extra-content-revenues/) (J. Clement, 2024a). Features like an engaging tale and well-developed characters are crucial to the experience of single-player games because they entice players to explore the game environment and advance the plot. But in GAAS games, the constant pursuit of gameplay precedes the story. Players will find it more difficult to relate to the game because they lack emotional investment in the characters and game environment. (Roso, 2024). Due to their accessible, free-to-play formats and constant updates, GAAS games have changed the gaming industry and drew gamers away from AAA titles. To satisfy player expectations, developers must now strike a balance between affordability, engagement, and storyline.

# Importance of the topic

Data Analytics contributes to many industries, and the game industry is no exception to this effect. As per a survey, in 2024, data analysts make up 4% of employees in the gaming industry in the United Kingdom (J, 2024). Game developers utilise analytics to make better games and players better at playing them. All developers want to create games that players enjoy, make more players play, and make more money for the publisher. Several attributes can indicate the success of a game, such as Daily Active Users (DAU), Monthly Active Users (MAU) and Average Revenue Per user (ARPU). For instance, as of Q3 2024, [Roblox](https://www.statista.com/statistics/1192573/daily-active-users-global-roblox/?ref=insights.geeiq.com) had around 71.5 million DAU(GEEIQ, 2024), and the ARPU for US mobile games is currently $57 per user, with a forecast of $64 by 2027 (Knezovic, 2024). With analytics, developers can improve the Key Performance Improvers (KPIs), guiding them on what to do next. Developers use analytics to help detect patterns in their user base and help them implement solutions to reduce churn or maximise playtime (Katy Yuan, 2024). Additionally, data science enables Dynamic Difficulty Adjustment (DDA), where the game automatically adjusts its difficulty based on the player's skill level. This ensures that players remain challenged and not frustrated, providing an optimal experience that encourages them to play the game (Rohan Whitehead, 2024).

The main objective of this proposed framework is to evaluate and predict the success of forthcoming PC games on the Steam platform by discovering the shared characteristics of successful and less successful titles, segmenting audiences through clustering for targeted marketing strategies, and accurately forecasting sales performance of upcoming AAA games. This will be accomplished using advanced data mining techniques and pre-processing methods: preparing the dataset, dealing with missing values, and classifying and standardising the outcome.

Accurately forecasting the success of forthcoming AAA games requires understanding and analysing AAA game data. Advanced analytics may be pivotal in using historical data to uncover key success elements and customise strategies for specific audience segments. By increasing forecast accuracy, developers and publishers may make better judgements to optimise marketing efforts and the potential success of new games.

# Research Objectives

Using data analysis, several research objectives can be put forward to predict video game sales. Three such **research objectives** are:

1. Clustering can be used to segment audiences and customise marketing efforts for each group. This can help discover the common traits of successful games, which can guide marketing and development strategies. (Reutterer and Dan, 2020; Sifa, Drachen and Bauckhage, 2021).
2. To predict the potential success or failure of upcoming PC games that are yet to be released over the Steam Platform using the data we collected previously from the same platform (Kerim and Genc, 2020; Pfau *et al.*, 2022).
3. To forecast the sales estimate of upcoming PC games that are yet to be released over the Steam Platform (Aziz *et al.*, 2018; Zhang *et al.*, 2019).

These research questions can accompany a comprehensive study of implementing and developing data-driven solutions for detecting, predicting, and forecasting by implementing (1) Clustering algorithms, (2) Classification algorithms, and (3) Regression analysis on the dataset.

# Data gathering and data analysis

## Feature Selection and Literature Review

Feature selection is an essential part of data gathering. It involves obtaining the most applicable attributes that can impact the accuracy of predicting the success of forthcoming games, reducing noise and other obstacles.

Steam is a well-known digital distribution platform, and research on user engagement, market trends, and gaming behaviour is highly relevant. The factors affecting player engagement, game success, and monetisation.

The data is collected based on several aspects, such as the information on [SteamDB](https://steamdb.info/), an unofficial website that provides detailed information and statistics about games available on Steam, and [games-stats.com](https://games-stats.com/), which provides details such as revenue and ratings unavailable over [SteamDB](https://steamdb.info/). A brief review of some of the essential attributes that have been studied and found to be effective in predicting the success of upcoming PC games:

* 1. Game’s Age (in days):(Johannes, Vuorre and Przybylski, 2021; Cerezo-Pizarro *et al.*, 2023) Age of the video game since the release.
  2. Price (in $): (Zhao and Ni, 2022) The game cost in USD.
  3. Game Followers: (De Luisa *et al.*, 2021) The number of followers for on Steam.
  4. All-Time Review: (Zuo, 2018; Al Mursyidy Fadhlurrahman *et al.*, 2023) Steam review converted to a scale of 1 to 7.
* 7 - Overwhelmingly positive
* 6 - Very Positive
* 5 - Mostly Positive
* 4 - Mixed
* 3 - Mostly Negative
* 2 - Very Negative
* 1 - Overwhelmingly Negative
  1. Game Reviews Count: (De Luisa *et al.*, 2021; Obedkov, 2021, 2022) The Steam review count.
  2. Net Revenue (in $): (Turner, 2022) Total revenue since release.
  3. User Rating (out of 10): (Al Mursyidy Fadhlurrahman *et al.*, 2023) User rated score based on review and feedback.

## Data Gathering

Data collection requires getting reliable data from several sources, evaluating it with various algorithms, and anticipating future outcomes. This approach is consistent with the research goals of leveraging data science and analytics to affect the gaming business. Here is a detailed outline of the steps involved:

* 1. Define the research objectives related to the prediction of video game sales.
  2. Determine the data types and attributes required to answer these objectives: Age, Price, Game Followers, All-Time Review, Game Reviews Count, Net Revenue, and User Rating.
  3. Identify and locate the data sources through public databases such as [SteamDB](https://steamdb.info/), surveys, blogs, and websites like [games-stats.com](https://games-stats.com/).
  4. Obtain necessary permission and access the relevant data by following ethical and legal standards.

This proposal mainly focuses on the public **Steam Games, Reviews, and Rankings** dataset. The dataset is downloaded from [Kaggle](https://www.kaggle.com/datasets/mohamedtarek01234/steam-games-reviews-and-rankings) (Tarek, 2024) and is available in the public domain. The [final dataset](https://www.icloud.com/iclouddrive/04bdPIq-AiCZMqWMYssrc6y1g#Assignment_-1_Rough_Draft) combines data from the Kaggle dataset and data from [games-stats.com](https://games-stats.com/) and is manually cross-verified with [SteamDB](https://steamdb.info/). It has nine attributes and 240 rows containing details about the Age, Price, Developer, Publisher, Game Followers, All-Time Review, and Game Reviews Count from [Kaggle](https://www.kaggle.com/datasets/mohamedtarek01234/steam-games-reviews-and-rankings). Net Revenue and User Rating data from [games-stats.com](https://games-stats.com/). Table 1 describes nine variables, corresponding descriptions, data types, data levels, metrics, and sources.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variables | Description | Data Type | Data Level | Metrics | Sources |
| Game’s Age (in days) | The age of the video game since release date | Ratio | Quantitative | Days | (Johannes, Vuorre and Przybylski, 2021; Cerezo-Pizarro *et al.*, 2023) |
| Price (in $) | The price of the video game | Ratio | Quantitative | Currency in $ | (Zhao and Ni, 2022) |
| Developer | The Developer of the video game | Nominal | Qualitative | Text | (Rizani, Khalid and Iida, 2023) |
| Publisher | The Publisher of the video game | Nominal | Qualitative | Text | (Rizani, Khalid and Iida, 2023) |
| Game Followers | The number of followers for a game on Steam | Ratio | Quantitative | Number | (De Luisa *et al.*, 2021) |
| All Time Review (in scale of 1 to 7) | The all-time review for the video game | Ordinal | Quantitative | In the scale of 1 to 7 | (Zuo, 2018; Al Mursyidy Fadhlurrahman *et al.*, 2023) |
| Game Reviews Count | The number of reviews for the video game | Ratio | Quantitative | Number | (De Luisa *et al.*, 2021; Obedkov, 2021, 2022) |
| Net Revenue (in $) | The overall revenue of the game since release date | Ratio | Quantitative | Currency in $ | (Turner, 2022) |
| User Rating (out of 10) | The rating for a game by the Steam | Ordinal | Quantitative | Score out of 10 | (Al Mursyidy Fadhlurrahman *et al.*, 2023) |

**Table 1**: Overview of Steam Games Dataset prediction

of the success of upcoming PC games.

## Data Analysis

### The Framework:

The proposed model, illustrated below as a diagram, shows the research framework and applies classification, clustering, and regression analysis techniques to derive insights from game data and produce results for the research objectives. The data analysis framework that will be used is CRISP-DM (Cross-Industry Standard Process for Data Mining).

A diagram of data processing

Description automatically generated

**Figure 1**: Proposed Model for prediction of the success

of upcoming PC games on the Steam platform Dataset.

### This framework describes a data analytics process for the Steam platform dataset. It begins with business and data comprehension and moves to preparatory activities, including cleaning, collection, and reduction. For modelling projects, clustering (K-Means, DBSCAN), classification (SVM, Naive Bayes), and regression (Ridge, Polynomial) are used. Findings are assessed and used to provide helpful information.

### Steps to perform CRISP-DM framework to produce results for the research objectives:

Step 1: Business Understanding

The first stage in addressing the objectives is to obtain business knowledge of the data in the dataset.

Step 2: Data Understanding

This step entails gathering the data and preliminary analysing the data level, type, variable, and time to comprehend its properties. It is part of finding trends, recognising problems with data quality, and obtaining knowledge that will affect data preparation and modelling.

Step 3: Data Preparation

This step comprises all tasks needed to produce clean and applicable data for the modelling step. Data preparation involves three steps, as mentioned below:

* **Data Cleaning and Validation:** Use strategies like binning or clustering to handle outliers and add or remove missing values to guarantee consistency.
* **Data Aggregation:** Normalise data by compiling important metrics and determining the links between variables using correlation analysis.
* **Data Reduction:** Using sampling or dimensionality reduction techniques**, r**educe the dataset size while keeping the information necessary for practical analysis.

Step 4: Modelling

This stage involves selecting and executing relevant algorithms to train a prediction model. The dataset, study goals, size, and complexity influence the optimal prediction model. To provide trustworthy results, it is advised to test many models, compare their accuracy and precision, display the findings, and perform cross-validation. The following methods are used to provide a solution to each of the research objectives.

|  |  |  |
| --- | --- | --- |
| Objective | Methods | Why (with Specific Steam Data Variables) |
| Clustering for Audience Segmentation- (Reutterer and Dan, 2020; Sifa, Drachen and Bauckhage, 2021) Segment audiences and customise marketing efforts for each group. Methods such as OPTICS can also be used. | K-Means - To create a set of equal, round clusters out of related data. | Fragment games by **Followers, Price, Reviews, and Revenue** to target marketing. |
| DBSCAN - A method for locating densely packed data sets, regardless of their irregular size or form. | Find clusters of highly engaged users using **Reviews, Followers, and Sentiment**. |
| Classification for Predicting Game Success/Failure - (Kerim and Genc, 2020; Pfau *et al.*, 2022) Predict whether a game will succeed or fail based on historical data. Methods such as KNN can also be used. | SVM – To clearly demarcate two groups, particularly in cases when the data is complicated. | Predict *Success/Failure* based on **Followers, Ratings, Reviews, and All-Time Reviews**. |
| Naïve Bayes – A quick and simple method for determining which category something belongs to when dealing with a lot of data. | Use simple models to forecast success with **Price, Reviews, and Followers.** |
| Regression for Forecasting Sales Estimates - (Aziz *et al.*, 2018; Zhang *et al.*, 2019) Forecast sales for upcoming games based on historical data. Methods such as Linear and Lasso regression can also be used. | Ridge Regression – For identifying trends in data and forecasting sales, especially in cases when several hints are tightly connected. | Estimate revenue using **Followers, Reviews, and Ratings** while avoiding data issues. |
| Polynomial Regression – For identifying a curve pattern in sales data, however it must be applied cautiously to prevent complications. | Study non-linear effects of **Ratings, Reviews, and Price** on sales without overcomplicating. |

**Table 2**: Overview of why the data analysis approach is selected.

Step 5: Evaluation

This step checks the findings to ensure they are valid and valuable.

Step 6: Deployment

In the last phase, a conclusion will be made when the deployment outcomes are presented.

# Conclusion

This study suggests a clustering structure to separate players and boost marketing efforts. The proposed model also sheds light on the traits of a successful game, allowing publishers and developers to steer the production process and stick to a marketing plan, increasing the likelihood that future games will be successful. Classification uses data analysis to forecast the results of upcoming PC games on the Steam Platform, giving creators a tool to help them make wise choices. Regression analysis enables businesses to make an accurate sales forecast, ensuring better planning and allocation of resources.

This framework has several drawbacks despite its potential. The approach may not apply to some platforms or game ecosystems like PlayStation, Xbox, etc., due to the use of data from Steam. (Sifa, Drachen and Bauckhage, 2021). Representativeness and forecast accuracy can be hampered by demographic bias, inconsistent data quality, and restricted access. These issues can be resolved by adding a variety of trustworthy datasets and increasing model flexibility. Notwithstanding these challenges, the research advances predictive analytics, facilitating improved business plans and data-driven game creation. It is a significant step toward enhancing decision-making and encouraging creativity in the gaming sector.

# Bibliography

Aziz, A. *et al.* (2018) ‘Empirical analysis on sales of video games: a data mining approach’, *Journal of Physics: Conference Series*, 1049, p. 12086. Available at: https://doi.org/10.1088/1742-6596/1049/1/012086.

Cerezo-Pizarro, M. *et al.* (2023) ‘The cultural impact of video games: A systematic review of the literature’, *Educ. Sci. (Basel)*, 13(11), p. 1116.

GEEIQ (2024) ‘How to measure success in gaming: 10 key gaming metrics you need to know - GEEIQ’, *How to measure success in gaming: 10 key gaming metrics you need to know - GEEIQ* [Preprint]. Available at: https://geeiq.com/how-to-measure-success-in-gaming-10-key-gaming-metrics-you-need-to-know/.

J, C. (2024) ‘UK games industry job role distribution 2021’, *Statista* [Preprint]. Available at: https://www.statista.com/statistics/1096364/job-role-distribution-in-the-games-industry-uk/.

J. Clement (2024a) ‘Gaming monetization - Statistics & Facts’, *Statista* [Preprint]. Available at: https://www.statista.com/topics/3436/gaming-monetization/.

J. Clement (2024b) *Topic: Video game industry*, *Statista*. Available at: https://www.statista.com/topics/868/video-games/.

Johannes, N., Vuorre, M. and Przybylski, A.K. (2021) ‘Video game play is positively correlated with well-being’, *R. Soc. Open Sci.*, 8(2), p. 202049.

Katy Yuan (2024) ‘Data analytics in gaming industry: a simple case study’, *MachEye* [Preprint]. Available at: https://www.macheye.com/blog/analytics-revolutionizing-gaming/.

Kerim, A. and Genc, B. (2020) ‘Mobile Games Success and Failure: Mining the Hidden Factors’, in *2020 7th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, pp. 167–171. Available at: https://doi.org/10.1109/ISCMI51676.2020.9311587.

Knezovic, A. (2024) ‘200+ mobile games statistics: market & revenue report [2024]’, *Udonis Mobile Marketing Agency* [Preprint]. Available at: https://www.blog.udonis.co/mobile-marketing/mobile-games/mobile-gaming-statistics.

De Luisa, A. *et al.* (2021) ‘Predicting the popularity of games on Steam’.

Al Mursyidy Fadhlurrahman, J. *et al.* (2023) ‘Sentiment Analysis of Game Reviews on STEAM using BERT, BiLSTM, and CRF’, in *2023 International Conference on Electrical Engineering and Informatics (ICEEI)*, pp. 1–6. Available at: https://doi.org/10.1109/ICEEI59426.2023.10346219.

Obedkov, E. (2021) ‘Simon Carless on how you can count game sales based on number of user reviews on Steam’, *Game World Observer* [Preprint]. Available at: https://gameworldobserver.com/2021/12/14/simon-carless-on-how-you-can-count-game-sales-based-on-number-of-user-reviews-on-steam.

Obedkov, E. (2022) ‘Steam reviews/sales ratio approaching 60x for new hit games, so how do we estimate sales in 2022?’, *Game World Observer* [Preprint]. Available at: https://gameworldobserver.com/2022/11/15/how-to-count-game-sales-steam-2022-review-multiplier.

Pfau, J. *et al.* (2022) ‘Predicting success factors of video game titles and companies’, in B. Göbl et al. (eds) *Entertainment Computing – ICEC 2022*. Cham: Springer International Publishing, pp. 269–282. Available at: https://doi.org/10.1007/978-3-031-20212-4\_22.

Reutterer, T. and Dan, D. (2020) ‘Cluster analysis in marketing research’, in C. Homburg, M. Klarmann, and A. Vomberg (eds) *Handbook of Market Research*. Cham: Springer International Publishing, pp. 1–29. Available at: https://doi.org/10.1007/978-3-319-05542-8\_11-2.

Rizani, M.N., Khalid, M.N.A. and Iida, H. (2023) ‘Application of meta-gaming concept to the publishing platform: Analysis of the Steam games platform’, *Information (Basel)*, 14(2), p. 110.

Rohan Whitehead (2024) ‘The role of data science in the gaming industry’, *The Role of Data Science in the Gaming Industry* [Preprint]. Available at: https://ioaglobal.org/blogdetails/role-of-data-science-in-gaming-industry/.

Roso, Z. (2024) ‘Quo vadis gaas gaming: the rise and fall of aaa titles amidst the single-player renaissance’. Available at: https://www.linkedin.com/pulse/quo-vadis-gaas-gaming-rise-fall-aaa-titles-amidst-zoran-vofre.

Sifa, R., Drachen, A. and Bauckhage, C. (2021) ‘Large-scale cross-game player behavior analysis on steam’, *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 11(1), pp. 198–204. Available at: https://doi.org/10.1609/aiide.v11i1.12804.

Tarek, M. (2024) ‘Steam games, reviews, and rankings.’, *Kaggle* [Preprint]. Available at: https://www.kaggle.com/datasets/mohamedtarek01234/steam-games-reviews-and-rankings.

Turner, A. (2022) ‘Video game industry revenue & market share(2024)’. Available at: https://www.bankmycell.com/blog/video-game-industry-revenue.

‘Video game’ (2024) *Wikipedia* [Preprint]. Available at: https://en.wikipedia.org/w/index.php?title=Video\_game&oldid=1259453817.

Zhang, C. *et al.* (2019) ‘Video Game Sales Analysis and Prediction Based on Machine Leanring’, *Video Game Sales Analysis and Prediction Based on Machine Leanring* [Preprint]. Available at: https://cs-4641-team-12.github.io/Project/.

Zhao, Y. and Ni, Y. (2022) ‘The pricing strategy of digital content resources based on a Stackelberg game’, *Sustainability*, 14(24), p. 16525.

Zuo, Z. (2018) ‘Sentiment analysis of steam review datasets using naive bayes and decision tree classifier’. Available at: https://hdl.handle.net/2142/100126.

# Appendices

The dataset forprediction of the success of upcoming PC games on the Steam platformis attached below and in Canvas as a Comma-Separated value (CSV) file:

[Steam Dataset\_(October\_25th\_2024)](https://www.icloud.com/iclouddrive/04bdPIq-AiCZMqWMYssrc6y1g#Assignment_-1_Rough_Draft)