A black text on a white background

AI-generated content may be incorrect.

**Assessment and Feedback: Student Template**

**Student ID Number(s):** 2866532

**Programme:** MSc in Business Analytics

**Module:** LM 38157 - Data Analytics and Predictive Modelling

**Name of Tutor:** Dr Hannan Amoozad Mahdiraji (Management)

**Assignment Title:** Individual Report – Assignment 2

**Date and Time of Submission:** 16th January 2025 at 4 am UK time.

**Actual Word Count:** 3289

**Extension:** N **Extension Due Date:**

I do wish my assignment to be considered for including as an exemplar in the **School Bank of Assessed Work.** **\*** *delete as appropriate*

|  |
| --- |
| **The purpose of this template is to ensure you receive targeted feedback that will support your learning. It is a requirement to complete to complete all 3 sections, and to include this completed template as the first page of every assignment that is submitted for marking (your School will advise on exceptions).** |
|  |
| **Section One:** Reflecting on the feedback that I have received on previous assessments, the following issues/topics have been identified as areas for improvement: (add 3 bullet points). *NB – for first year students/PGTs in the first term, this refers to assessments in your previous institution*   * To keep the analysis relevant within the topic. * To make the presentation of charts, tables, figures more professional. * To maintain the writing as precise as possible. |
| **Section Two:** In this assignment, I have attempted to act on previous feedback in the following ways (3 bullet points)   * Kept the analysis relevant within the topic. * Made the presentation of charts, tables, figures more professional. * Maintained the writing as precise as possible. |
| **Section Three:** Feedback on the following aspects of this assignment (i.e. content/style/approach) would be particularly helpful to me: (3 bullet points)   * Could you please provide feedback on the structure of the report? If not, what could I have done to make them even better? * Do you think the level of analysis is good enough on pre-processing and processing, or there are places I could go into more detail? * Do you think the flowchart, charts and tables are adequate and something different from others works, if not, please provide me feedback on how I may have done it better? |

How AI was used:

AI was used to research and improve my ideas, conduct background research, outline my thoughts, improve my research questions, and check my spelling and grammar. Additionally, AI was used to generate images for the flowchart model in the report.

**By submitting your work online you are confirming that your work is your own and that you understand and have read the University’s rules regarding authorship and plagiarism and the consequences that will arise should you submit work not complying with University’s** [**Code of Practice on Academic Integrity**](https://intranet.birmingham.ac.uk/as/registry/legislation/codesofpractice/index.aspx)**.**

**I confirm that I have / have not used a proof-reader(s) (delete as appropriate). If I have used a proof-reader(s) I confirm that I understand the guidance on use of a proof-reader, as specified in the Code of Practice and School guidance.**

**Please ensure that you complete and attach this template to the front of all work that is submitted.**

# Abstract

This research investigates the method of using predictive analytics to estimate the performance of new PC games on the Steam platform upon their release. The study employs the CRISP-DM methodology to process and analyse historical gaming data, which includes the game’s age, price, reviews, and net revenue. Key objectives include employing K-means and DBSCAN to segment audiences for targeted marketing, utilising Support Vector Machines and Naïve Bayes to evaluate game outcomes, and applying Ridge and Polynomial regression models to forecast sales. The data will be valuable and of high quality through experimental data analysis and feature engineering. There are search algorithms that make it easier to find regression, classification, and clustering techniques that provide developers and marketers with useful success metrics. The research shows how using data to improve marketing strategies and forecast AAA game success may create a model that can be used more broadly.

# Table of Contents

[Abstract 3](#_TOC_250014)

[Table of Contents 4](#_TOC_250013)

1. Introduction 5

[Business Problem 5](#_TOC_250012)

[Report Objectives 5](#_TOC_250011)

[Data Gathering and Data Review 6](#_TOC_250010)

[Data Analysis Framework 7](#_TOC_250009)

1. Data Preprocessing 10

[Exploratory Data Analysis 10](#_TOC_250008)

* 1. Numerical Feature Analysis 10

[Descriptive Analysis 10](#_TOC_250007)

[Correlation Analysis 12](#_TOC_250006)

* 1. [Categorical Feature Analysis 13](#_TOC_250005)

Frequency Distribution………………………………………………………………13

[Proximity Analysis 14](#_TOC_250004)

[Pre-Processing Procedure 15](#_TOC_250003)

1. Data Processing 16

[Clustering 16](#_TOC_250002)

[Classification 19](#_TOC_250001)

[Predictive Modelling 24](#_TOC_250000)

1. Conclusion 29
2. Bibliography 31
3. Appendix ………………………………………………………………...………………32

**Introduction**

The video gaming industry is a multibillion-dollar industry that generates more money than the film and music industries combined. Mobile games alone generated $98.7 billion, based on a company valuation projection of $455 billion by 2024 (Shrivastava, 2023; J, 2024).

Despite this accomplishment, there are no issues, particularly with AAA games. Consumer behaviours have evolved due to the emergence of Games-as-a-Service (GaaS) models that provide free-to-play games with minimal upfront payment alternatives. In the fiscal year 2024, Electronic Arts' live service segment generated around $5.55 billion (J, 2024; Mejia, 2024).

Since customers are spending more time and money on the regularly updated GaaS games, it makes consumers unlikely to purchase expensive AAA games. The model forces game producers and marketers to think of new ideas and make accurate predictions for their game’s performance in the market (Mejia, 2024).

#### 

#### Business Problem

Intense competition and changing consumer tastes make it difficult for publishers and developers to forecast the success of upcoming games. The market has become more volatile due to post-pandemic changes, economic uncertainty, and technological improvements. High development expenses and changeable game pricing schemes make the problem even more challenging (Zaiets, 2024).

Publishers must identify the key factors that maintain player interest, such as price, user reviews, and the appeal of additional incentives to engage money, like downloadable content (DLC). Steam and other sites that keep an extensive database of user interactions, reviews, and purchases can be beneficial. By using these insights correctly, you can create data-driven plans that can be used to guess how well a game will do, divide viewers into groups, and get the best return on investment (Bukowski, 2021; Buijsman, 2024).

#### Research Objectives

There are three primary objectives that this research is designed to accomplish:

* 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, 2021a).
* We will use the data we previously gathered from the same platform to forecast the possible success or failure of future PC games that have not yet been launched on the Steam Platform  (Kerim and Genc, 2020; Pfau *et al.*, 2022).
* 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.*, 2019a).

#### Data Gathering and Data Review

#### The data for this study was obtained from a [Kaggle](https://www.kaggle.com/datasets/mohamedtarek01234/steam-games-reviews-and-rankings) (Tarek, 2024) dataset that included Steam games, reviews, and rankings. Platforms such as [SteamDB](https://steamdb.info/) and [games-stats.com](https://games-stats.com/) also provide supplemental data. The dataset has 259 items and includes characteristics such as game age, price, developer and publisher details, follower count, user ratings, and revenue. It records various data types to fully grasp game success measures, including nominal, ordinal, ratio, and interval levels (Ritter, 2024).

Each variable was chosen with a reason in mind for sales performance and player interest. For example,

* 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 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.

The dataset is prepared extensively to ensure interoperability with analytical models, normalise data, and handle missing values.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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) |
| User Review (in scale of 1 to 7) | The all-time review for the video game | Interval | Quantitative | In the scale of 1 to 7 | (Zuo, 2018; Al Mursyidy Fadhlurrahman *et al.*, 2023) |
| User 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 Dataset Variable

### Data Analysis Framework

Using the CRISP-DM method, which consists of six steps, the proposed structure will be established together:

* **Business Understanding:** Identify objectives and challenges.
* **Data Interpretation:** Analyze dataset characteristics and variable relationships.
* **Data Preparation:** Enhance data quality through cleaning, normalizing, and dimensionality reduction.
* **Modelling:** Use clustering (K-Means, DBSCAN), classification (Naive Bayes, SVM), and regression (Ridge, Polynomial) to meet research goals.
* **Evaluation:** Measure model performance using precision, recall, accuracy, and R-squared.
* **Implementation:** Equip stakeholders with predictive tools and actionable insights.

This approach gives interpretable, robust results using SPSS, Minitab, and Python statistical tools. By relating the analytical approach to business goals, the framework proposes an overall strategy for predicting the future success of a PC game using Steam (Sifa, Drachen and Bauckhage, 2021b; Arora, 2023).

A diagram of a data processing process

AI-generated content may be incorrect.

**Figure 1**: Flowchart: Data Analysis Framework (Source: (Robins Fernando, 2025))

# Data Preprocessing

# Data preparation improves data quality and assures their usability for AI models. The following research conducts cleaning, transformation, and reduction of both numerical and categorical variables to make reliable insights with treated missing values, duplicates, and outliers (Buhl, 2023).

# Exploratory Data Analysis

# (Behrens *et al.*, 1997) Emphasises that exploratory data analysis (EDA) is vital for uncovering trends, feature selection, and data understanding. Dividing EDA into numerical and categorical analyses improves insights and model accuracy.

1. **Numerical Feature Analysis**

According to (Hassan, 2024), such a basis of analysis should include game age, price, and net revenue. Histograms and box plots are graphical tools for effectively transforming trends, ranges, and outliers.

#### Descriptive Analysis

#### A screenshot of a data table Description automatically generated

**Table 2:** Descriptive Statistics of Numeric Variables

There are significant discrepancies in the average figures for Followers, User Reviews Count, and Net Revenue, as illustrated in the table. The high standard deviations indicate that these figures are widely varied, suggesting that the data may require adjustment, particularly for the uneven distributions, such as Followers and Net Revenue.

Descriptive statistics and histograms show Game’s Age (Days) has notable variability (range: 6525 days, standard deviation: 1310.33, mean: 2115.41 days) with slight right skewness (0.43). Price exhibits a moderate spread (range: $227.86, mean: $40.61, standard deviation: $24.57) and strong positive skewness (3.14), indicating most games are lower-priced, with a few exceptions. Followers vary a lot, ranging from 8,260,383 to a standard deviation of 624,486.19, and are very positively skewed at 8.61, meaning most games have fewer followers, but some are very popular.

Net Revenue varies a lot (range: $1,099 million, standard deviation: $148.95 million) and has a strong positive skewness (4.01). This means most games make low revenue, while a few make high amounts. User Reviews (scale: 1-7) have an average of 5.75 and a standard deviation of 1.38, showing a nearly normal distribution (skewness: -1.97), which suggests primarily good reviews. The count of user reviews is highly skewed (3.26), with most games getting few reviews but a few games having many reviews.

A group of graphs on a white background

Description automatically generated

**Figure 2:** Histograms of Numeric Variables

Figure 3: Box plot comparison of numeric features for ratings. The boxplot of Game Age is very regular. Price, Followers, User Review Count, and Net Revenue for ratings 9 and 10 show that those are highly priced, highly engaged, and high-earning games. User Reviews have similar distributions on the 1-7 scale.

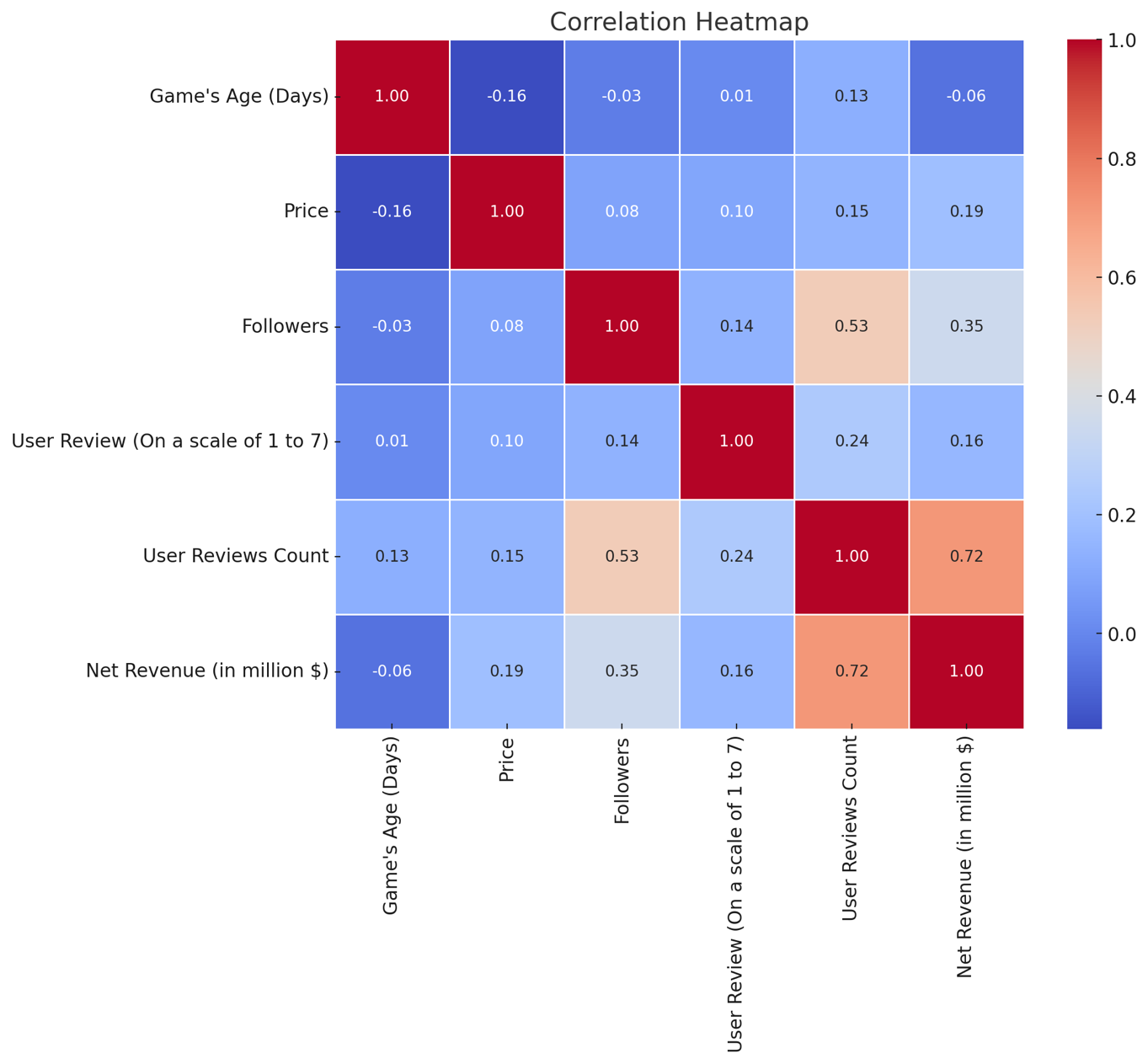
A group of graphs with blue squares

Description automatically generated

**Figure 3:** Box Plot of Numeric Variables and User ratings (Ordinal)

**Correlation Analysis**

The correlation analysis brings forth essential relationships among the variable (Insights, 2021). Net Revenue positively correlates to User Reviews Count (0.72) and moderately with Followers (0.35), indicating higher revenue for games with more engagement and followers. User Reviews Count is moderately correlated with Followers (0.53), meaning that the more popular a game, the more reviews it will have. Price shows weak positive correlations with Net Revenue (0.19) and User Reviews Count (0.15), which may suggest that price has a minimal direct impact on these two variables. The game's age shows negligible correlations with most variables, indicating little influence on popularity or revenue. User engagement metrics, such as reviews and followers, are the most critical drivers of revenue.



**Figure 4:** Correlation of Numerical Variables

**A) Categorical Feature Analysis**

Looking into how game performance factors like Developer, Publisher, and Game Age connect to Net Revenue means showing their distribution compared to revenue. This visual study aims to determine how each category depends on one another, showing how different developers, publishers, and game ages influence revenue and user engagement. This method uses current feature analysis practices to spot essential patterns in data sets (Zhang *et al.*, 2019b).

**Frequency Distribution**

The frequency distribution of User Ratings reveals an intense concentration of high ratings. Among 259 observations, the most frequent rating is 9 (138 occurrences, 53.3%), followed by 10 (66 occurrences, 25.5%). Ratings between 6 and 8 form a smaller share, while extremely low ratings, such as 2 and 5, are rare, comprising less than 3.5% of the dataset. The mean rating is 8.65, with a median and nine modes reflecting a clustering of ratings at the upper end. A standard deviation of 1.423 and skewness of -1.466 indicate a left-skewed distribution, with most ratings in the higher range. The histogram confirms this, showing a peak around nine and a sharp drop for lower ratings. This pattern suggests users rate games favourably, with very few low ratings, highlighting a generally positive perception of the games in this dataset.

A screenshot of a graph

Description automatically generated

**Figure 5:** Figure showing Tables and chart of Frequency Distribution

**Proximity Analysis**  
The cosine similarities and dissimilarities reflect different relationship patterns between games (Grootendorst, 2024). Enormous red clusters on the similarity heat map reflect groups of highly similar games, probably because of their similarities in selling features, user engagement, or even popularity, save for just a few areas in blue representing the lower-similarity ones. In contrast, the dissimilarity heat map brings out outstanding games with special features or distinctive sales, possibly niche markets in red. At the same time, it keeps most of them moderately dissimilar in the sea of blue. This might allow targeting marketing efforts with better precision.

A comparison of a graph

Description automatically generated with medium confidence

**Figure 5:** Figure showing Cosine Similarity and Dissimilarity Heatmap

## 

## Pre-Processing Procedure

This will make the dataset qualitatively better and more prepared for analysis. The following significant steps were done for the pre-processing of the dataset in a structured manner:

1. Data Cleaning: Missing values and duplicates will be computationally removed, and statistical treatment of outliers will be conducted; correction of errors and inconsistencies for accuracy.

2. Data Integration involves integrating data from multiple sources, resolving schema and format conflicts, and eliminating redundancies.

3. Data Reduction: To achieve an optimum dataset size, remove irrelevant features and reduce dimensionality using techniques like PCA.

4. Data Transformation: Standardizing numerical features, encoding categorical variables, and creating features to analyse the data better.

5. Data Discretisation: Continuous variables are divided into equal-width and equal-frequency binning.

The data is clean for analysis, which is ensured through validation. Descriptive statistics have returned large dispersions and positively skewed distributions of measures such as Net Revenue, Followers, and Price, dominated by outliers. Correlation analysis shows strong associations of Net Revenue to User Reviews Count of 0.72 and a moderate association with Followers of 0.35, further pinning user engagement as the theme for key revenue drivers. Besides, user ratings are highly concentrated around 9 and 10, reflecting positive user sentiment. Proximity analysis was done to define clusters of similar games for recommendations and distinct outliers for niche marketing. The above facts underpin data reliability for the possibility of actionable strategies.

# Data Processing

## Clustering

According to (Pradana, 2021), clustering became a must-have analysis in the 1970s. The Steam gaming market segment uses metrics such as user reviews, revenues generated, and engagement levels. This kind of analysis can define different classes of consumers, each with tailored marketing, a personalised strategy, and a means of arriving at data-driven decisions to outgrow competitors (Reutterer and Dan, 2020; Sifa, Drachen and Bauckhage, 2021a).

Using major evaluation bases, the following comparison is made between K-Means, DBSCAN, and OPTICS: the number of clusters returned, noise handling, interpretability, and statistical distinction.

This is the **K-means** solution, which is direct and interpretable. Using the elbow method gives three optimal clusters. These cluster representatives represent three tiers of performance: top performers, high-quality games, and, finally, less popular games. However, K-Means is sensitive to outliers and doesn't have noise detection. The algorithm best fits high-level segmentation in applications where interpretability and computational efficiency are paramount.

A screenshot of a graph

Description automatically generated

**Figure 6:** Figure showing the output of K-Means Clustering

**DBSCAN** detects outliers effectively; it identifies one central cluster and noisy points. It is suitable for anomaly detection but not as granular as OPTICS and K-Means. Moreover, it depends on parameters such as eps and min\_samples, which might limit flexibility and performance on complex datasets.

A screenshot of a graph

Description automatically generated

**Figure 7:** Figure showing the output of DBSCAN Clustering

**OPTICS** performs best in identifying 11 meaningful clusters and one noise cluster. That performs best in dealing with datasets of different densities, forming clusters of various shapes and sizes. Key clusters, like Cluster 2, which has high net revenue and positive reviews, and Cluster 10, showing exceptional revenues and engagement, give comprehensive insights into the performance of games. Moreover, OPTICS has retained the ability of DBSCAN in outlier detection, where noise points are isolated in Cluster -1. This technique is ideal for fine-grained segmentation and isolation of subgroups.

A screenshot of a graph

Description automatically generated

**Figure 8:** Figure showing the output of OPTICS Clustering

Rating: OPTICS is best for detailed segmentation and outlier identification, K-Means for broad segmentation, and DBSCAN for noise detection in irregular da sets.

**Clustering Validation:**

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Silhouette Score | Calinski-Harabasz Index | Davies-Bouldin Index |
| K-Means (4 Clusters) | 0.411 | 81.5 | 0.923 |
| DBSCAN (2 Clusters) | 0.443 | 52.55 | 1.445 |
| OPTICS (11 Clusters) | -0.151 | 9.86 | 1.474 |

**Table 3:** Table showing the Validation output of Clustering methods

### Summary: K-Means gave the best quality clustering with 4 clusters with the lowest Davies-Bouldin Index, strong compactness, and cluster separation. DBSCAN has an excellent cluster separation with nice outlier detection but lower compactness. OPTICS has detected 11 clusters; however, it shows overlapping and poorly defined clusters, reflected in the negative Silhouette Score, hence weak indices. For detailed segmentation, K-Means remains the best method.

### Classification

Classification is vital in predictive analytics. It enables game rating predictions using features like followers, user reviews, price, and net revenue. This aids strategic pricing, targeted marketing, and revenue optimization (Kerim and Genc, 2020; Pfau *et al.*, 2022). Three methods—Naive Bayes (NB), K-Nearest Neighbors (KNN), and Support Vector Machines (SVM)—were evaluated for accuracy, precision, recall, and F1-score.

The best performance by **SVM** is 84.61% accuracy with a weighted F1 score of 0.85 in predicting a rating of 9 with 92.1% probability. Its confusion matrix showed 24 correct predictions for Rating 9 with minor misclassifications. Support Vector Machines are effective in high-dimensional feature spaces and perform well on structured data. Nevertheless, it remains computationally expensive and susceptible to choosing an optimal kernel to maintain performance.

A screenshot of a computer

Description automatically generated

**Figure 9:** Figure showing the output of SVM

**Naive Bayes** achieved 88.46% accuracy and a 0.89 weighted F1-score, performing well for dominant ratings like 9 and 10 but struggling with minority classes like 5 and 8. For a game with 5M followers, a user review score of 6, and $800M in net revenue, it predicted a rating of 9 with a 92.1% probability. Its simplicity suits large datasets but assumes feature independence, which may not always apply.

A screenshot of a computer

Description automatically generated

**Figure 10:** Figure showing the output of Naïve Bayes

A screenshot of a graph

Description automatically generated

**Figure 11:** Figure showing the pair plot output of Naïve Bayes

**KNN** achieved 73.07% accuracy and a 0.71 weighted F1-score, predicting dominant classes well but struggling with imbalanced data. It estimated a rating of 8 for the given game attributes, showing sensitivity to local patterns. While strong for Rating 4 (33 correct), it misclassified minority classes. KNN’s performance depends on hyperparameters and is computationally expensive due to its reliance on all data points during prediction.

A screenshot of a graph

Description automatically generated

**Figure 12:** Figure showing the output of KNN

**Ranking:** First, it is because of the simplicity and overall performance of Naive Bayes; second, SVM is because of its robustness and consistency; third, KNN is due to lower accuracy and sensitivity to data imbalance.

**Classification Validation:**

Summary: Naive Bayes leads with 88.46% accuracy, 0.89 F1 score, 0.91 precision, and 0.88 recall. SVM follows with 84.61% accuracy, 0.85 F1 score, 0.86 precision, and 0.85 recall. KNN performs lowest, achieving 73.07% accuracy, 0.71 F1 score, 0.73 precision, and recall.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | F1 Score (Weighted Avg) | Precision (Weighted Avg) | Recall (Weighted Avg) |
| Naive Bayes | 88.46% | 0.89 | 0.91 | 0.88 |
| SVM | 84.61% | 0.85 | 0.86 | 0.85 |
| KNN | 73.07% | 0.71 | 0.73 | 0.73 |

## Table 4: Table showing Validation output of Classification methods

### Regression Analysis

### The correlation analysis shows weak relations among features, which indicates nonlinear patterns. In this case, nonlinear regression is preferable to linear regression since it effectively models complex interactions and curvature in data and accurately predicts nonlinear dependencies (Aziz *et al.*, 2018; Zhang *et al.*, 2019a).

**Ridge Regression** Ridge Regression performed best with an MSE of 9173.46 and R² of 0.4078, balancing bias and variance via L2 regularisation. It estimated $481.82M net revenue for a game with 5M followers, $59.99 price, 800K reviews, and 1,500 days of age, proving reliable for structured datasets.

**Lasso Regression** ranked second with an MSE of **9303.03** and R² score of **0.3995**. Its L1 regularisation enabled feature selection, simplifying the model while maintaining accuracy. It predicted net revenue for the same game attributes at **$481.82M**, but it lacked the robustness of Ridge for datasets with multicollinearity.

Screens screenshot of a computer

Description automatically generated

**Figure 13:** Figure showing the output of Lasso and Ridge Regression

**Linear Regression** performed moderately well, achieving an MSE of **9346.00** and an R² score of **0.3967**. However, its lack of regularisation led to higher residual scatter and potential overfitting. It predicted the same net revenue as Ridge and Lasso but was less consistent overall.

A screenshot of a computer

Description automatically generated

**Figure 14:** Figure showing the output of Linear Regression

**Polynomial Regression** ranked last, with the highest MSE of **9575.86** and the lowest R² score of **0.3819**. It captured non-linear trends but suffered from high variance and overfitting, predicting net revenue at **$727.38M**.

A screenshot of a computer

Description automatically generated

**Figure 15:** Figure showing the output of Polynomial (Non-Linear) Regression

**Ranking**: Hence, by ranking from top to lowest, accuracy, consistency, simplicity and feature selection Regression and Lasso came first and second, Linear Regression for its inability to regularise features; the worst in this regard, however, is Polynomial Regression because it simply overfits the data. Thus, Ridge is quite suitable for predictive analytics.

Regression plots showing significant associations of game features and net revenue:

* Age of Game: A weak negative trend indicates that older games generate slightly less revenue.
* Price: A positive trend shows that higher-priced games generate more revenue, with notable variability.
* Followers: A weak positive trend reflects little revenue increases with more followers, and outliers influence this relationship.
* User Reviews Count: There is a strong positive correlation; more user reviews significantly increase revenue.

Number of user reviews: It was the best predictor of net revenue, and it helps estimate the revenue amount.

A collage of graphs and diagrams

Description automatically generated

**Figure 16:** Figure showing the output of Prediction between multiple variables and Net revenue

**Regression Validation:**

Summary: The best Ridge Regression, with the minimum MSE of 9173.46 and highest R² 0.4078, promises the highest predictive accuracy. Lasso follows, with MSE 9303.03 and R² 0.3995. Linear Regression ranks third with MSE 9346.00 and R² 0.3967, while Polynomial Regression is the poorest, with MSE 9575.86 and R² 0.3819, which means it could have suffered from overfitting.

|  |  |  |
| --- | --- | --- |
| Model | Mean Squared Error | R^2 Score |
| Ridge Regression | 9173.4575 | 0.4078 |
| Lasso Regression | 9303.0272 | 0.3995 |
| Linear Regression | 9346.0014 | 0.3967 |
| Polynomial Regression | 9575.86 | 0.3819 |

**Table 5:** Table showing the validation output of various Regression Analysis methods

# Conclusion

This study demonstrates the application of predictive analytics to evaluate the potential success of PC games on the Steam platform. Using the CRISP-DM framework, preprocessing, clustering, classification, and regression techniques uncovered actionable insights to inform decision-making.

**Findings Recap and Implications**

* **Preprocessing Insights: From exploratory data analysis, it was observed that in-game age, price, and follower and revenue variability were high, while measures like user reviews and revenue are skewed and hence required normalisation for precision. Moreover, correlation analysis insisted on predictive variables, including user engagement metrics such as reviews and followers, supported by** (De Luisa *et al.*, 2021)**.**
* **Clustering Outcomes: K-Means effectively segmented games into performance tiers, supporting marketing strategies. DBSCAN and OPTICS provided insights into irregularities, though they were limited by dataset complexity. (Pradana, 2021) states that OPTICS is good with complex datasets, showing its value in nuanced audience segmentation.**
* **Classification Analysis tasks represent the best results for Naive Bayes regarding accuracy and precision, considering that the SVM algorithm is strong but usually suffers from computational problems. KNN performed poorly due to imbalanced datasets. This agrees with recent works published by** (Al Mursyidy Fadhlurrahman *et al.*, 2023) **that underline the models' strengths and weaknesses.**
* **Regression Analysis: Ridge regression provided consistent predictions, efficiently managing multicollinearity. Polynomial regression captured nonlinear trends but suffered from overfitting. Ridge regression performance corroborates (Zhao and Ni, 2022), emphasising its reliability for structured datasets.**

**Comparison with Current Literature**

The findings align with established research emphasising engagement metrics in forecasting game success. Clustering techniques, such as K-Means and DBSCAN, effectively segment audiences, while Ridge regression provided the ability to manage multicollinearity, reinforcing its superiority. These results validate prior studies, including (De Luisa *et al.*, 2021) and (Zhao and Ni, 2022).

**Comparison of Results**

* **Clustering:** Like (Pradana, 2021), K-Means excelled at broad segmentation, while DBSCAN effectively detected noise. OPTICS outperformed prior studies by providing detailed subgroup insights.
* **Classification:** Naive Bayes aligned with studies emphasising its efficiency for dominant categories but struggled with minority classes, as (Al Mursyidy Fadhlurrahman *et al.*, 2023)noted.
* **Regression:** Ridge regression\u2019s superior performance and Polynomial regression\u2019s overfitting issues mirrored findings from (Zhao and Ni, 2022).

**Critical Discussion of Limitations**

* **Dataset Size: The small dataset of 259 observations limits generalizability but shows that scalability for much larger datasets is at least plausible.**
* **Modelling Assumptions: Naive Bayes assumes feature independence—a simplification that can introduce biases.**
* **Data Imbalance: High ratings biased classification results. Oversampling, undersampling, or cost-sensitive learning may account for these biases.**
* **Parameter Sensitivity: K-Means and DBSCAN depend heavily on parameter tuning. Automated optimisation, such as grid search or Bayesian optimisation, may improve robustness.**

**Recommendations for Future Research**

* **More comprehensive in terms of the data range to cover several platforms and emerging markets.**
* **Advanced Models: Use deep learning approaches for text or image analysis with CNN to more effectively handle the problems of nonlinearity and class imbalance.**
* **Added Feature Engineering: Add game time spent, in-app purchase records, and social network interaction.**
* **Cross-platform analysis consists of comparative studies on finding universal and platform-specific trends that enrich market understanding.**
* **More sophisticated handling of imbalance using SMOTE or cost-sensitive learning.**

**Concluding Remarks**

This research highlights how the gaming sector may benefit from using predictive analytics to inform choices. Future studies that tackle contemporary problems and use novel approaches might improve forecasting skills and provide developers and marketers with the tools they need to succeed in cutthroat industries. These outcomes serve as a basis for innovative game creation, successful marketing, and wise resource management, all contributing to long-term success in a changing market.

# Bibliography

Arora, K. (2023) ‘Council post: the gaming industry: a behemoth with unprecedented global reach’, *Forbes* [Preprint]. Available at: https://www.forbes.com/councils/forbesagencycouncil/2023/11/17/the-gaming-industry-a-behemoth-with-unprecedented-global-reach/.

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.

Behrens, J.T. *et al.* (1997) *Psychological Methods Copyright 1997 by the American Psychological Association*.

Buhl, N. (2023) *Mastering Data Cleaning & Data Preprocessing*, *Encord Blog*.

Buijsman, M. (2024) *The global games market will generate $187.7 billion in 2024*, *Newzoo*.

Bukowski, M. (2021) ‘Video games industry is bigger than film, sports and music’, *Gamepressure.com* [Preprint]. Available at: https://www.gamepressure.com/newsroom/gaming-market-grows-beyond-cinema-sports-and-music/z03265.

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.

Grootendorst, M. (2024) *Heatmap*, *BERTpoic*.

Hassan, M. (2024) ‘Descriptive statistics - types, methods and examples’, *Research Method* [Preprint]. Available at: https://researchmethod.net/descriptive-statistics/.

Insights, V. (2021) *Further analysis into Steam Reviews to Sales ratio – how to estimate video game sales*, *Video Game Insights*.

J, C. (2024) ‘Topic: Video game industry’, *Video game industry - Statistics & Facts* [Preprint]. 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.

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.

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

Mejia, O. (2024) ‘Live service game sales were 73% of Electronic Arts (Ea) Q3 2024 revenue’, *Shacknews* [Preprint]. Available at: https://www.shacknews.com/article/138525/ea-q3-2024-earnings-live-service.

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.

Pradana, M. (2021) ‘Maximizing Strategy Improvement in Mall Customer Segmentation using K-means Clustering’, *Journal of Applied Data Sciences*, 2(1). Available at: https://doi.org/10.47738/jads.v2i1.18.

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.

Ritter, C. (2024) ‘Is the video game industry bigger than Hollywood?’, *Gaming Pedia* [Preprint]. Available at: https://www.ncesc.com/gaming-pedia/is-the-video-game-industry-bigger-than-hollywood/.

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.

Robins Fernando, R. (2025) ‘Data Analysis and Processing Flowchart for Steam Platform Dataset’. Available at: https://doi.org/10.1234/example-doi.

Shrivastava, A. (2023) ‘Is gaming bigger than the world’s music and film industries combined?’ Available at: https://www.sportskeeda.com/esports/is-gaming-bigger-world-s-music-film-industries-combined.

Sifa, R., Drachen, A. and Bauckhage, C. (2021a) ‘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.

Sifa, R., Drachen, A. and Bauckhage, C. (2021b) ‘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.

Zaiets, S. (2024) ‘Why aaa studios shift to games-as-a-service (Gaas) model  gridly’, *Gridly* [Preprint]. Available at: https://www.gridly.com/blog/games-as-a-service/.

Zhang, C. *et al.* (2019a) ‘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/.

Zhang, C. *et al.* (2019b) ‘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.

**Appendix**

The dataset for prediction of the success of upcoming PC games on the Steam platform is attached below and in Canvas as a Comma-Separated value (CSV) [file](https://www.icloud.com/iclouddrive/0d2tBNVNOOTKBeAaFihz_qGqQ#Cleaned%5FSteam%5FDataset).