Project Final report

Project Title: What Makes a Video Game Successful?

Project Final Report: Video Game Sales Analysis

Course: STA 4241 - Statistical Learning

Name : Munchootsorn Wangsriviroj, William Rocha, Jackson Windhorst, Sean Murray, Nicholas Clewley

Project Goal

The goal of this project is to understand what factors affect video game sales. Using real-world data, this report explores trends and builds simple models to help predict global sales based on features like genre, platform, and ratings. We are aiming to help and aid those game studios and developers to align with what the trends seem to follow according to our model that we will create. It will consist of many variables that can influence those global sales that can be influential for game sales. This project will dissect the landscape of video game sales by constructing a simple but effective model.

Executive Summary

This report explores what makes a video game successful using a dataset from Kaggle with over 16,000 records. The analysis focuses on how factors such as genre, platform, critic and user scores relate to global sales. The data was cleaned, visualized, and used in predictive modeling. Linear regression and decision tree models were tested, and clustering methods grouped games based on performance. Results show that timing of release and critic scores have a strong impact on global sales. This insight can support better planning and decision-making in the gaming industry overall.

Project Final report

Model Building Process

1. **Define the Problem**

   The objective is to predict global video game sales and identify what makes a video game successful.
2. **Data Collection**

   The dataset was collected from Kaggle and includes over 16,000 records of video game sales. It contains information such as global and regional sales, genre, platform, user and critic scores, and developer.
3. **Exploratory Data Analysis (EDA)**

   See Section 1.6 below.
4. **Modeling**

   Models include Linear Regression, Lasso Regression (for feature selection), Decision Trees, and Clustering methods such as K-Means and Hierarchical Clustering.
5. **Evaluation**

   Model results were evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2).
6. **Deployment**

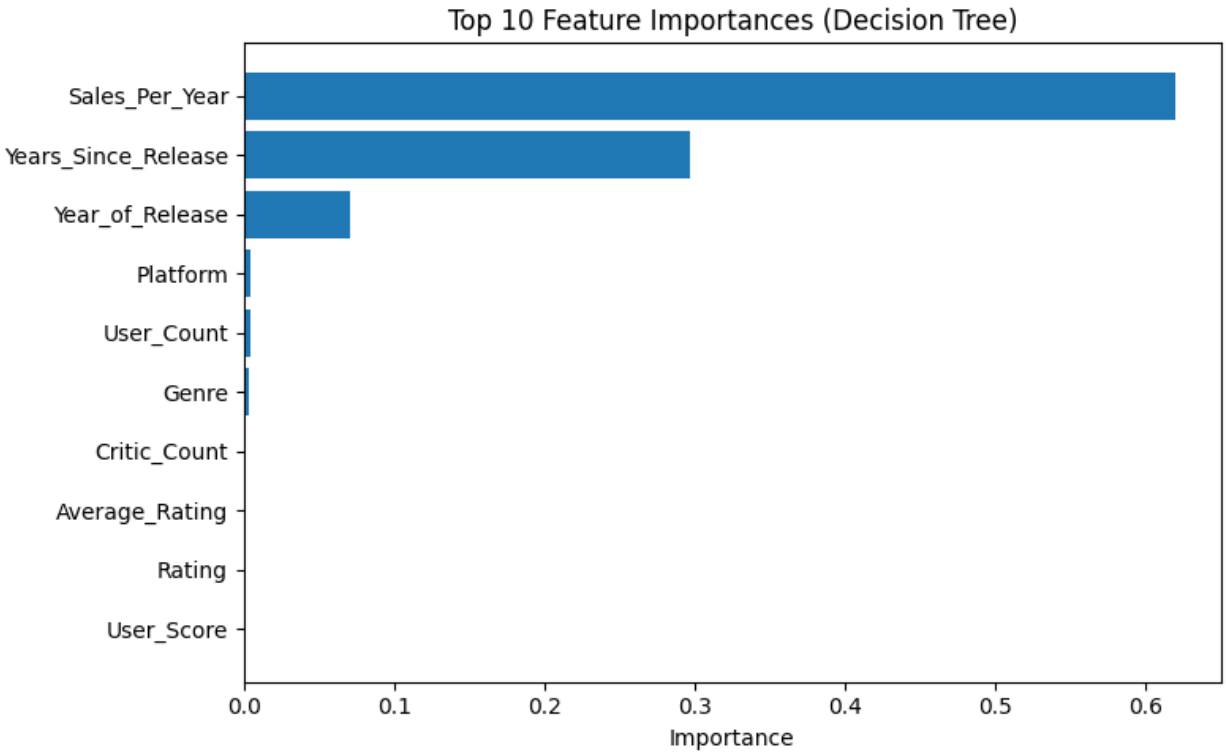
   The models developed in this project have been deployed and the following results from them can be useful for game publishers and developers to predict their games' sales, improve planning for future projects, and better strategize new game releases.

1.2 Model Evaluation

Model	MAE	RMSE	R^2
Linear Regression	0.403	1.445	0.504
Decision Tree	0.202	1.139	0.692
Lasso Regression	0.202	1.445	0.5041

Project Final report

Based on the metrics, the Linear regression model appears to be performing well with a R-squared of 0.504, meaning that 50% of the variance in global sales based on the features we are using: platform, year of release, genre, publisher, critic score, critic count, user score, user count, rating, years since release, sales per year, and average rating. With the lasso regression, it has a similar r-squared of 0.5041 using only 6 factors: Sales per year, publisher, critic score, critic count, years since release, and user count. However this model has a lower MAE showing that it has a better average absolute difference between the prediction and the actual global sales of 0.202. Moving to the Decision tree, it appears to have the strongest and best performing model based on the metrics: R-squared of 0.692. It also has the lowest MAE and RMSE, suggesting it provides the most accurate prediction and explains the most variance in global sales among the three models.

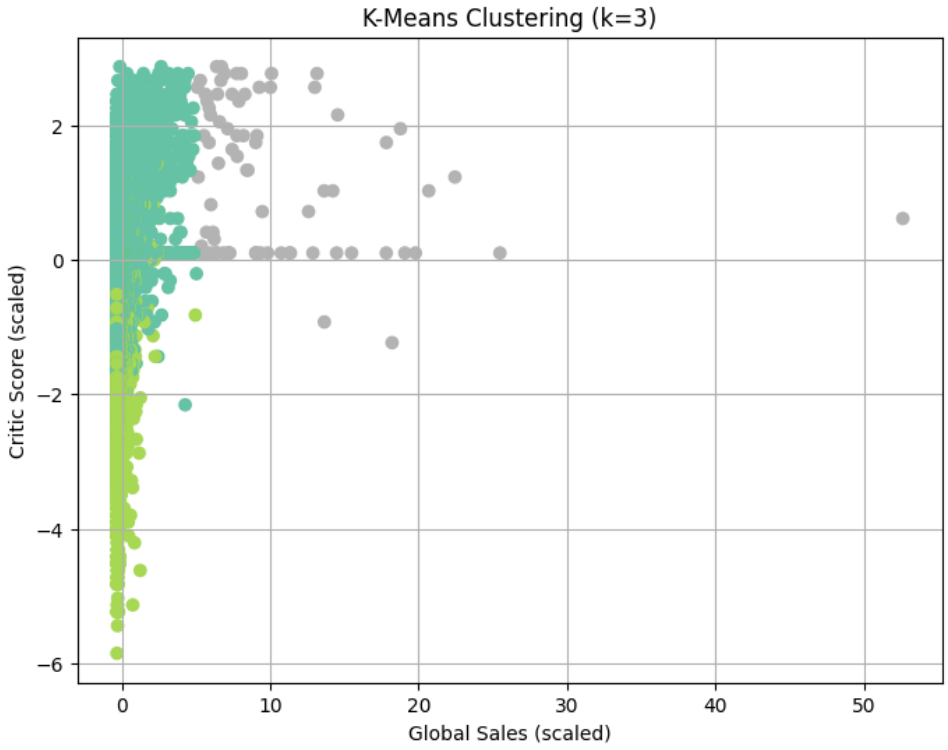


Project Final report

Clustering:

- **K-Means Clustering (K=3)**

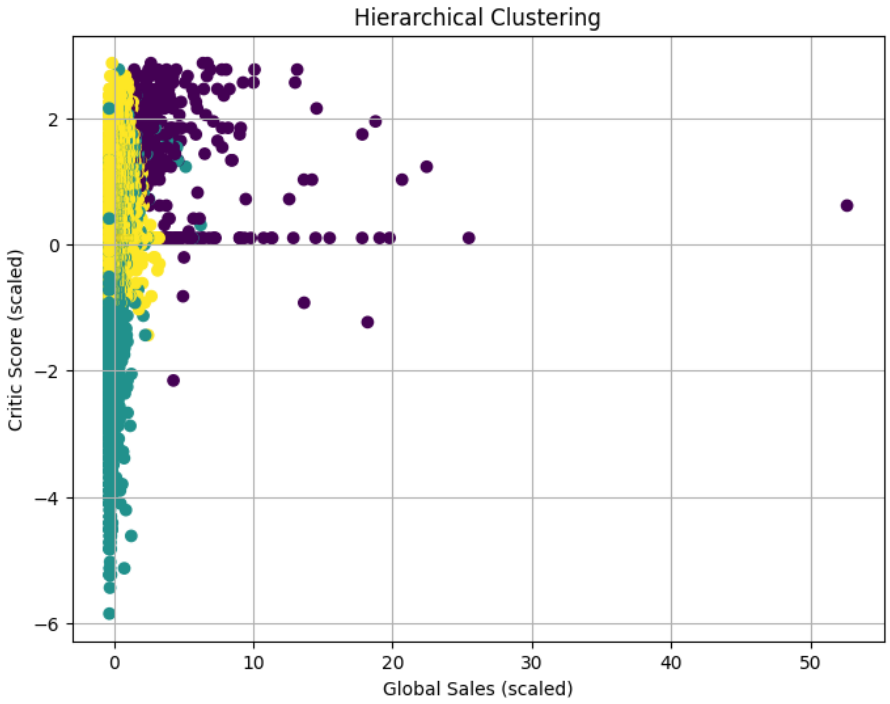
Project Final report



- We grouped the games into 3 clusters based on Global Sales, Critic Score, and USer Score. In cluster 0, there were 14,249 games which indicate the games with the lowest sales explaining that many common games don't shine or go viral in the industry of gaming development. In cluster 1, there were 2,120 games showing the games that have high critic scores but still didn't shine in sales. These games show that games that have high critics don't necessarily predict global sales. In cluster 2, there are only 79 games that look like outliers but are the ones with the highest sales and scores. Even though these games were grouped, it doesn't help us to determine a pattern or model for our big question.

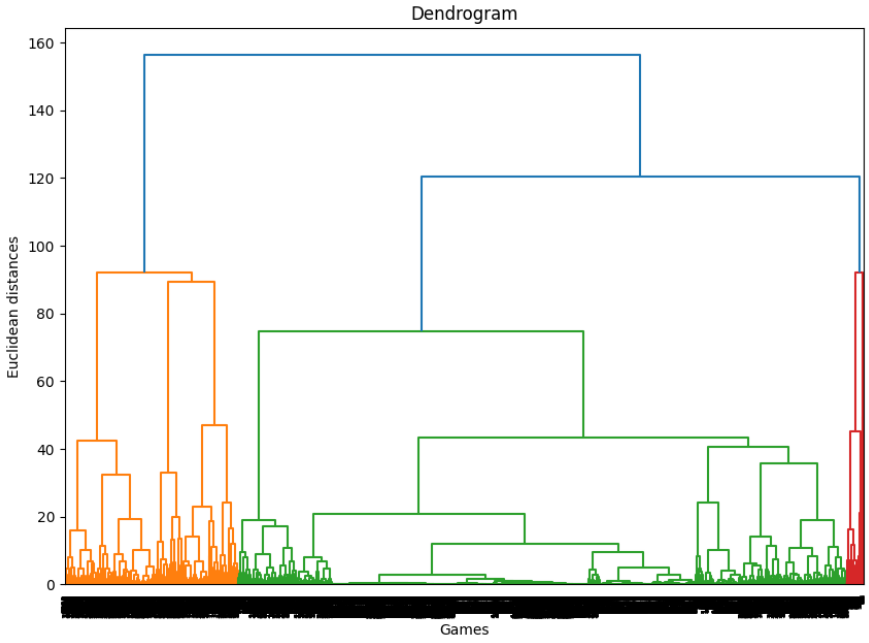
Project Final report

- **Hierarchical Clustering**



  - To gain further insight into any more clusters we used hierarchical to obtain better clusters. Since we implemented k means before hierarchical and obtained decent

results this step became almost irrelevant and was a chance to implement just another unsupervised learning model.



1.3 Learning Algorithms Implemented

- **Linear Regression:** Used to predict sales based on features such as platform, critic score, and user count.
- **Lasso Regression:** Helped select only the most important features.
- **Decision Tree Regression:** Showed which variables had the highest importance.
- **K-Means Clustering:** Grouped games with similar scores and sales into clusters (output in appendix).
- **Hierarchical Clustering:** Grouped games using distance metrics and showed results in a dendrogram (in appendix).

1.4 Reproducibility

A .Zip file is submitted with the Jupyter Notebook, which includes all code, charts, and outputs. Running the notebook will generate all results in this report.

Project Final report

1.5 Executive Summary for Stakeholders

The analysis shows that global game sales are affected by features such as platform, region, and scores from critics and users. Models were used to explain and predict sales patterns. These results can help video game companies understand trends and focus on important features during development and release planning.

1.6 Exploratory Data Analysis (EDA)

The dataset was cleaned by removing rows with missing names or release years. Missing publisher and developer values were replaced with “Unknown.” Non-numeric user scores were handled and converted correctly.

EDA Results:

- **Top platforms:** PlayStation 2 had the highest total sales.
- **Sales trend:** Game sales peaked around 2008.
- **Genres by region:** Different regions preferred different game genres.
- **Ratings comparison:** Critic and user ratings showed a similar trend but they did not always agree.
- **Top developer:** Nintendo had the highest total sales among developers.

A heatmap showed strong correlation between regional sales (especially NA and EU) and global sales. Additional charts showed trends over time and differences in genre popularity.

Project Final report

1.7 Final Thoughts

Through this project, we set out to better understand what drives video game sales, and our results offer valuable insight for both developers and publishers. By analyzing a large and diverse dataset, we identified key factors such as: critic score, release timing, and publisher influence. Across our modeled timeframe, these factors consistently correlate with higher global sales so we have to assume they'll continue to do so. Our decision tree model stood out as the most impressive in its predictive power and can be easily interpreted by business people to be used as a guiding strategy for future game releases.

While our models are simple by design, they still capture meaningful patterns that game studios can use to improve planning and positioning. There's room for future improvement by using more recent data, additional marketing variables, or more advanced modeling, but our work demonstrates a strong baseline foundation for how data-driven decisions can provide a real, practical edge in an extremely competitive industry.

In the end, success in gaming isn't random. It follows trends, preferences, and measurable factors. With the right data and the right questions, we can bring clarity to what makes a hit and that's exactly what we accomplished here.

References:

- Dataset: https://www.kaggle.com/datasets/xtyscut/video-games-sales-as-at-22-dec-2016csv

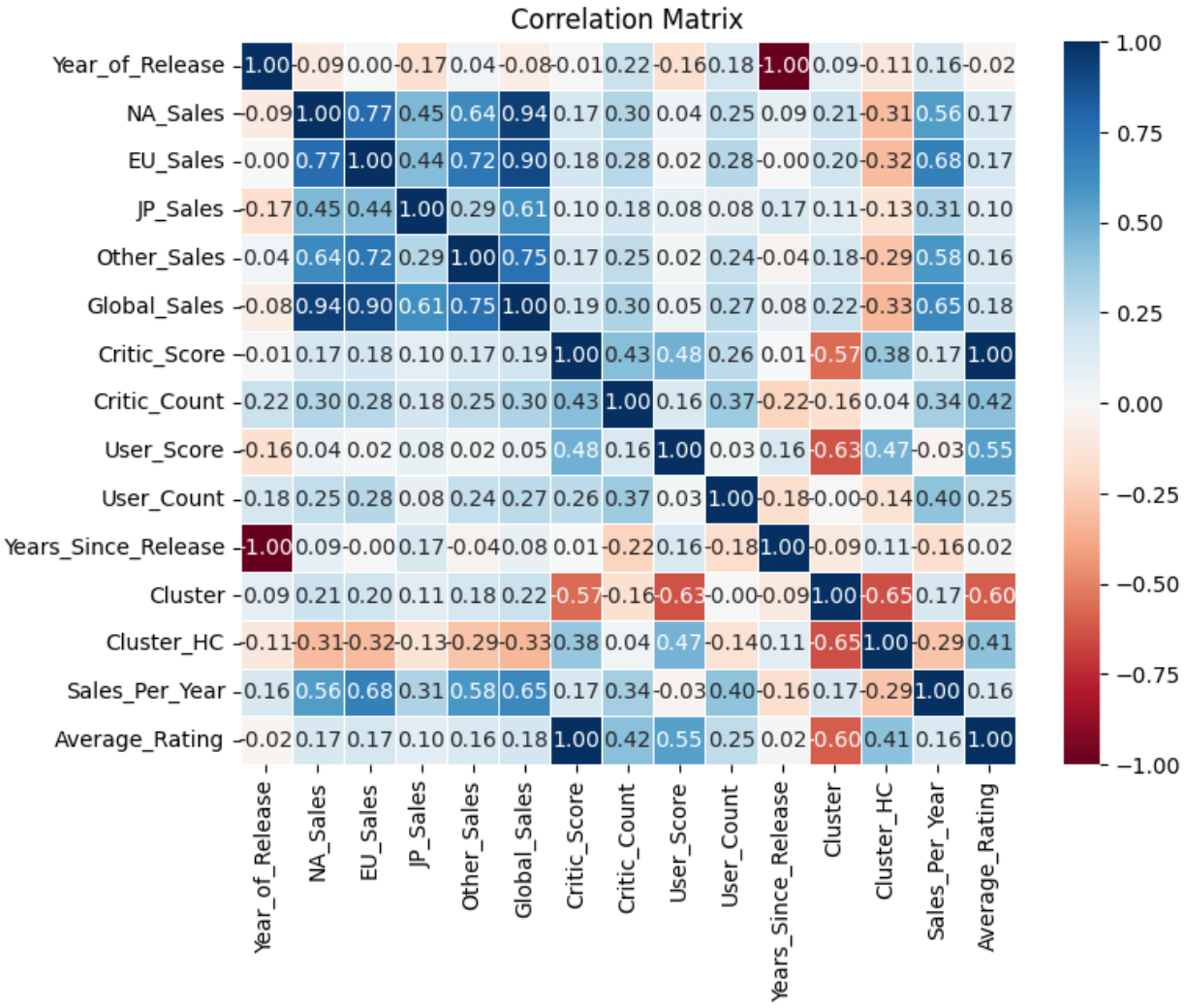
- James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An introduction to statistical learning: With applications in Python. Springer.
- McKinney, W. (2022). Python for Data Analysis: Data wrangling with pandas, NumPy, and Jupyter. O'Reilly Media, Inc.
- McKinney, W. (2022). Python for Data Analysis: Data wrangling with pandas, NumPy, and Jupyter. O'Reilly Media, Inc.
- Wang, W. (2025). Principles of Machine Learning: The Three perspectives. Springer.

Project Final report

Appendix : Visualizations

Project Final report

Heat Map:



Exploratory Data Analysis Visualizations

Project Final report