# DS 3000 – Dataset

# Project Topic Idea: <Game Sale Prediction*>*

1. Problem Statement: Because of the prosperity of the game industry, more and more games come into the market per year. This project focus on the sale of games on Steam. We will analyze the factor affecting the price of a game. We might collect data of previous sales, cost, game genre, developer’s name, rate and amount of new game publication. Data from game releasing platforms, social media and related datasets is needed in this process.

2. Significance of the Problem: For players or people working in related fields, sale prediction gives people a simple idea of the game and its company. For game companies, sale prediction is necessary because the prediction gives a guide to their future plan. This project not only predict the sale of the game, it also indicates the reason, which is significant for the game company to determine their next step: managing adverting budget, scheduling publication time of the game, or whether to make new games of one series.

3. Potential Datasets Potential datasets:

• Kaggle www.kaggle.com/tamber/steam-video-games/data www.kaggle.com/gregorut/videogamesales [www.kaggle.com/rush4ratio/video-game-sales-with-ratings](http://www.kaggle.com/rush4ratio/video-game-sales-with-ratings)

(this is the one we used for FP2 dataset)

• Steam Dataset steam.internet.byu.edu/

• Portsmouth Research Portal researchportal.port.ac.uk/portal/en/datasets/video-games-dataset(d4fe28cd-1e44-4d2f-9db6- 85b347bf761e).html Online information: Twitter / Instagram Use APIs to get posts. Use key words to grab information.

## Dataset File

Download or scrape your data from the source you identified above. Save your dataset as a CSV file. The first row of the file should contain variable names.

Describe your variables below (add more rows if necessary):

<https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings>

|  |  |  |
| --- | --- | --- |
| **Variable name in file** | **Description** | **Feature/ Outcome** |
| Name | Name of the game | feature |
| Platform | Gaming platform | feature |
| YearofRelease |  | feature |
| Genre |  | feature |
| Publisher |  | feature |
| NASales |  | Outcome |
| EUSales |  | Outcome |
| JPSales |  | Outcome |
| OtherSales |  | Outcome |
| Global\_Sales |  | Outcome |
| Critic\_score | Aggregate score compiled by Metacritic staff | feature |
| Criticcount | The number of critics used in coming up with the Criticscore | feature |
| User\_score | Score by Metacritic's subscribers | feature |
| Usercount | Number of users who gave the userscore | feature |
| Developer | Party responsible for creating the game | feature |
| Rating | The ESRB ratings | feature |
| In the Feature/Outcome column, indicate whether the variable is a feature or outcome variable. | | |

**Based on what we discussed regarding machine learning, does your dataset include a set of feature variables and one outcome variable that you can use for a supervised machine learning task?**

**We did include a set of feature variables, and we have multiple sales as outcome variables. We would probably pick global sale, which is most relevant to the feature variables we have.**