# Dataset Name

Digital Humanities Project: Effects of Storytelling in Video Game Ratings and Sales

# Description

Originally published on Zenodo (http://doi.org/10.5281/zenodo.2454579), this dataset is part of a project made for a Master’s level university course. Data used for this project was harvested from Metacritic in 201 and cleaned to ensure statistical value. This dataset features 7000 rows and was normalized for upload into the GG! Repository.

#Variables

title

Title of video game

(string)

year

Year of the title’s release

(numeric value: YYYY)

genre

The genre or genres associated with the title

(string)

publisher

Company responsible for the title’s publication

(string)

naSales

Sales data from North America

(in millions)

euSales

Sales data from Europe

(in millions)

jpSales

Sales data from Japan

(in millions)

otherSales

Sales data from other geographic regions

(in millions)

globalSales

Total global sales data

(in millions)

criticScore

Title’s aggregate review score as compiled by Metacritic critics

(numeric value)

criticCount

Number of critics whose reviews contributed to the final score

(numeric value)

userScore

Title’s aggregate review score as compiled by Metacritic users

(numeric value)

userCount

Number of users whose reviews contributed to the final score

(numeric value)

developer

Company responsible for developing the title

(string)

esrbRating

The official ESRB rating for the title

(string)

# Normalization Notes

Normalization included transforming variable names to suit the repository’s preferred naming conventions.

The variables “Story Focus” and “Gameplay Focus” were removed from the dataset. Due to their subjectivity and the admitted uncertainty on the part of the dataset creator, they were deemed irrelevant to this repository’s purpose.

The variable “Series” was also removed. In the raw dataset, “Series” specifically refers to “a collection of games that follow the same character… exist within the same world… or feature character development elements between them.” This definition suits the original project but does not align with the repository’s broader definition of a series. This discrepancy combined with the checkmark format of the variable’s associated values (in which “x” indicates a title is a series while a null value indicates it is not), this variable was also deemed irrelevant.

# Usage and Limitations

This dataset can be used to identify video game titles with the most recorded sales and highest Metacritic review scores as of 2016. It is possible to titles titles with more sales or higher scores compared to others released in the same year, in the same genre, by the same publisher, by the same developer, or within the same ESRB rating category.

While it is possible to identify titles with the most sales in a specific geographic region, this dataset only includes data specific to North America, Europe, and Japan. Other regions are rolled into “otherSales,” limiting the dataset’s usage for evaluating sales in Africa, Oceania, South America, or Asia at large.

As the number of video game players increase worldwide every year, older video games are at a disadvantage for sales. To compensate for this bias, this dataset is not filtered for games that have sold more than a specific number of units.

# Authors and Acknowledgement

Timo Poutanen and the University of Helsinki

# License

Attribution 4.0 International (CC BY 4.0)