Video Games Dataset

Team-2

Members: Tarlan Sultanov, Manasi Patil, Mansi Sawant, Zain Ahmed (1959323), Enes Tanrikulu, Ninad Dhatrak

Introduction:

In the gaming industry, data is generated from multiple sources such as sales, player reviews, and game ratings. This project focuses on Web Data Integration to combine disparate datasets into a unified structure for comprehensive analysis. By integrating data from various platforms, we can derive insights related to game sales, player engagement, and market trends, empowering stakeholders like developers, publishers, and marketers.

Our goal is to integrate several datasets related to video games into a single dataset that covers key attributes such as game name, platform, developer, release year, sales, and player engagement. This enables analysis of game performance and player sentiment in a unified framework.

Dataset description:

| **Dataset Name** | **Source** | **Format** | **# of Entities** | **# of Attributes** | **List of Attributes** |
| --- | --- | --- | --- | --- | --- |
| Video Game Sales Dataset | Kaggle | CSV | 16,598 | 10 | Name, Platform, Year of Release, Genre, Publisher, Global Sales, NA Sales, EU Sales, JP Sales |
| Steam Store Games Dataset | Kaggle | CSV | 27,075 | 8 | Name, Release Date, Developer, Publisher, Genres, Categories, Price, Tags |
| Video Games Dataset | Kaggle | CSV | 12,483 | 9 | Name, Platform, Year of Release, Publisher, Genre, Developer, ESRB Rating, Critic Score, User Score |

Dataset Compliance with Requirements:

| **Dataset Name** | **Requirement** | **Compliance Status** | **Evidence** |
| --- | --- | --- | --- |
| Video Game Sales Dataset | Completeness > 90% | Met | All key attributes are > 98% complete |
| Steam Store Games Dataset | Attribute Diversity | Met | Includes attributes such as Price, Genres, Publisher |
| Video Games Dataset | User Ratings & Critic Scores | Met | ESRB Rating and both Critic and User Scores present |

Descriptive Statistics of Key Attributes:

| **Dataset Name** | **Attribute** | **Mean** | **Median** | **Std Dev** | **Missing Values (%)** |
| --- | --- | --- | --- | --- | --- |
| Video Game Sales Dataset | Global Sales | 0.55 | 0.23 | 1.3 | 2% |
| Video Games Dataset | User Score | 7.3 | 7.5 | 1.2 | 0% |
| Steam Store Games Dataset | Price | 10.75 | 5 | 15.4 | 5% |

1. Use Case Description

The use case of this project is to integrate three distinct datasets related to video games to create a comprehensive database that facilitates insights into the video gaming industry. By combining data from various sources, we can gain a deeper understanding of market trends, game performance, and consumer preferences.

The project will unfold in three phases. Phase 1 involves standardizing the datasets for consistency. Phase 2 focuses on identity resolution, utilizing matching algorithms to eliminate duplicates. Phase 3 encompasses data fusion, aggregating data to generate reliable metrics like average sales and ratings. The expected outcomes include improved market insights and analysis of game sales trends, empowering stakeholders to make data-driven decisions and enhancing strategies in the gaming industry.

1. Datasets Requirements

Table 1. Datasets

| Dataset Name | Source | Format | # of entities | # of attributes | List of attributes |
| --- | --- | --- | --- | --- | --- |
| Video Game Sales Dataset | [Kaggle](https://www.kaggle.com/datasets/gregorut/videogamesales) | CSV | 16,598 | 10 | Name, Platform, Year of Release, Genre, Publisher, Global Sales, NA Sales, EU Sales, JP Sales |
| Steam Store Games Dataset | [Kaggle](https://www.kaggle.com/datasets/nikdavis/steam-store-games) | CSV | 27,075 | 8 | Name, Release Date, Developer, Publisher, Genres, Categories, Price, Tags |
| Video Games Dataset | [Kaggle](https://www.kaggle.com/datasets/mohamedhanyyy/video-games) | CSV | 12,483 | 9 | Name, Platform, Year of Release, Publisher, Genre, Developer, ESRB Rating, Critic Score, User Score |

Table 2. Attribute Intersection with Integrated Schema

| Attribute Name | Attribute Type | Dataset in which the attribute is found |
| --- | --- | --- |
| Name | String | Video Game Sales Dataset, Steam Store Games Dataset, Video Games Dataset |
| Platform | String | Video Game Sales Dataset, Video Games Dataset |
| Year of Release | Integer | Video Game Sales Dataset, Video Games Dataset |
| Genre | String | Video Game Sales Dataset, Steam Store Games Dataset |
| Publisher | String | Video Game Sales Dataset, Steam Store Games Dataset, Video Games Dataset |
| Developer | String | Steam Store Games Dataset, Video Games Dataset |
| Global Sales | Float | Video Game Sales Dataset |
| NA Sales | Float | Video Game Sales Dataset |
| EU Sales | Float | Video Game Sales Dataset |
| JP Sales | Float | Video Game Sales Dataset |
| Release Date | Date | Steam Store Games Dataset |
| ESRB Rating | String | Video Game Sales Dataset |
| Critic Score | Float | Video Game Sales Dataset |
| User Score | Float | Video Game Sales Dataset |
| Categories | List | Steam Store Games Dataset |
| Tags | List | Steam Store Games Dataset |

1. The selected datasets contain enough overlapping entities, such as common game titles, platforms, developers, and publishers, to ensure meaningful integration. Popular games like *The Witcher 3* and *FIFA* appear in multiple datasets, enabling cross-platform analysis. Shared attributes like *Game Name*, *Platform*, *Developer*, and *Publisher* will allow for schema matching and entity resolution, ensuring the integration of sales, pricing, and user engagement data across the datasets. This overlap supports the project’s goal of creating a unified view of game performance and player sentiment.