Kevin Danh

DSC 540 Data Preparation

Professor Williams

Oct 5, 2025

Project Milestone 1

This project aims to analyze trends in the video game industry by integrating multiple datasets and data sources. The analysis will explore factors such as sales, performance, awards, and ratings to uncover insights about the global video game market.

* Data Sources:
  + Flat File – Kaggle Dataset:
    - This contains global video game sales and performance data from 1980 to 2016.
    - <https://www.kaggle.com/datasets/modiash/video-game-sales-and-performance-data>
  + API – RAWG.io :
    - This provides a comprehensive video game data including descriptions, ratings, playtime, genres, platforms, and more.
    - <https://rawg.io/apidocs>
  + Website – Wikipedia Table:
    - This contains lists of Game of the Year awards given to various games across different regions and/or criteria.
    - <https://en.wikipedia.org/wiki/List_of_Game_of_the_Year_awards>
* Relationships
  + By using the name, genre, platform, and possibly publisher as the common key columns, there should be a relationship between these three datasets to explore such as:
    - How sales and performance metrics relate to awards and critical reception.
    - Trends across genres, platforms, and publishers over time.
    - Patterns in ratings, playtime, and global popularity.

Project Approach and Plan

The project will focus on analyzing trends in the video game industry by integrating multiple data sources, including a Kaggle dataset on global video game sales, the RAWG.io API for detailed game metadata, and Wikipedia tables listing games that achieve awards. The approach involves first cleaning and standardizing each dataset to ensure consistent naming conventions, genres, platforms, and publishers. Next, datasets will be merged using common key columns such as game name, genre, platform, and publisher. The integrated data will then be analyzed to identify trends in sales, ratings, and awards, as well as correlations between critical acclaim and commercial success. Visualizations such as time-series plots or bar charts can be used to highlight patterns across different regions, genres, and platforms.

Concerns and Challenges

Challenges may arise due to inconsistencies in naming conventions, duplicate entries, and missing or incomplete data across sources. There may be confusion in platforms for many games because it could be released on various platforms at once like PC and XBOX. Some games may have a later release date due to platform exclusivity such as God of War (2018) being a PS4 exclusive but then released for PC in 2022. Differences in platform or publisher naming, as well as variations in award titles, may require careful preprocessing. API rate limits and missing metadata from RAWG.io could restrict the completeness of the dataset. Also, some data may be limited in the year range such as one dataset having data until 2016 while the other has 2024 data.

Ethical Implications

While the project uses publicly available data, ethical considerations should be included such as respecting data usage policies and ensuring no copyrighted material is improperly used. Analysis should avoid drawing biased conclusions about specific demographics or developers/companies/publishers. Finally, visualizations and reports should accurately represent trends without exaggerating commercial or cultural significance while maintaining integrity and transparency within this analysis.