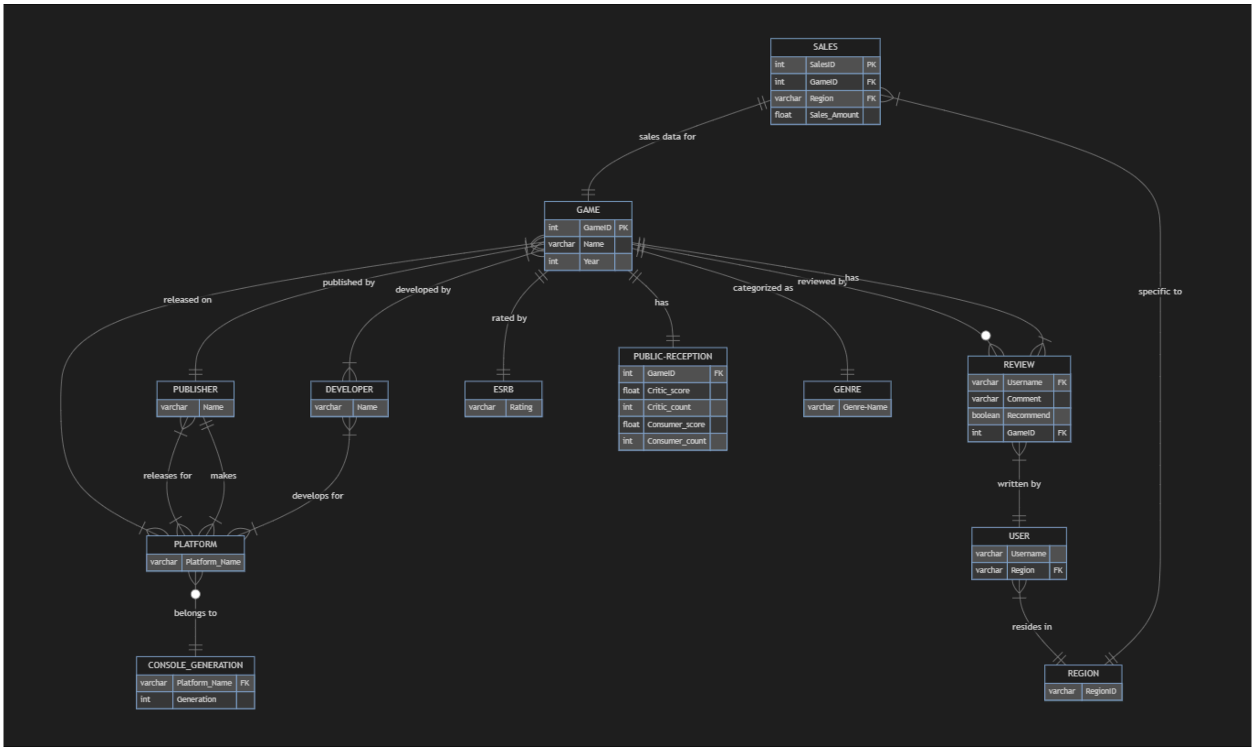
Michael Martinelli

Dataset:
https://www.kaggle.com/datasets/ibriiee/video-games-sales-dataset-2022-updated-extra-feat
Original Source: https://data.world/sumitrock/video-games-sales
ERD Diagram



Entities and Attributes

- Game
  - GameID, Name, Year
- Platform
  - Platform Name
- Publisher
  - Publisher Name
- Developer
  - Developer Name
- Sales
  - SalesID
  - GameId
  - Region
  - Sales Amount
- ESRB

    - Rating
    - GameId
- Public Reception
    - GameID
    - Critic_score
    - Critic_count
    - Consumer_score (alias for User_Score)
    - Consumer_count (alias for User_Count)
- Genre
    - Genre Name
- Region
    - RegionID
- Console_Generation
    - (New data, added to categorize various platforms in the dataset)
    - Platform Name
    - Generation
- User
    - Username
    - Region
- Review
    - Username
    - Comment
    - Recommend (yes or no boolean if they recommend to a friend)
    - GameId

Relations:

1. Game to Platform (Many to Many)
    a. Each game is released on **at least one** or more platforms.
2. Game to Publisher (Many to one)
    a. Each game is published by exactly one publisher. Publishers have multiple games that they have published.
3. Game to Developer (many to many)
    a. Each game is developed by **at least one** developer. Some games can have multiple development teams.While typically there is usually one main developer, you cannot exclude credited support studios.
4. Game to ESRB (one to one)
    a. Each game has **at most one** rating. Not every game has received a rating from the ESRB.
5. Game to Public Reception (one to one)

    a. Each game has at most one set of data on how the game scored to the critics and general public.Not game has set data on this

6. Game to Genre (one to one)
    a. Each game has **exactly one** genre
7. Game to Review (one to many)
    a. Each game can be reviewed by many users. Not every game has a review yet.
8. Sales to Game (one to one)
    a. Each sales data set is specific to exactly one game. There is multiple sales data based on region in this set, but each set of specific sales links back to one game.
9. Sales to Region (many to one)
    a. Each sales entry is specific to exactly one region. Not every game is sold in every region but is sold in at least one region.
10. Developer to Platform (many to many)
    a. Developers can create games for multiple platforms. Some developers are platform exclusive or publisher exclusive but regardless. Developers create games for at least one platform.
11. Publisher to Platform (many to many)
    a. Publishers can release games on multiple platforms. Some developers are platform exclusive.
12. Publisher to Platform (one to many)
    a. Some publishers release multiple platforms This is specific in the case of Microsoft, Sony, and Nintendo (but includes future publishers) that design and create their own Platforms.
13. Platform to Console Generation (many to one)
    a. Each platform is categorized into **at most one** console generation. (Exclusion of the PC since it is not a console). Many platforms belong to a single generation.
14. User to Region (many to one)
    a. Each user belongs to **exactly one** region but regions can include many users.
15. Review to User (many to one)
    a. Each review is written by exactly one user, while a user can write multiple reviews.
16. Review to Game (many to one)
    a. Each review is written for exactly one game, but games can have multiple reviews.

Extra Relations to consider:

- Genre to Platform: certain genres are more popular on specific platforms

- Genre to Region to Sales: certain genres sold better in specific regions
- ESRB to Genre: certain genres tend to relate to a specific ESRB score

Use cases for the dataset:

- Predicting future market trends based on history.
- Can analyze the market trends of specific genres in specific regions, comparing its competitors in the same console generation.
- Inventory management system for games that are in stock, and deciding which games to purchase for a store
- Analysis of gaming demographics and their preferred genres/platforms
- Analysis of competitors
- Impact of critic reviews
- Analysis for growth of the gaming industry and future predictions
- A system for users to leave feedback and reviews.
- Predict potential sales of a game that you might want to develop and what platform it will perform best on.

Topic Choice: Datasets and gathering new data for Machine learning

In the world of computer science, everyone is discussing the future of AI and the ethics behind it. Everyday there are new advancements in artificial intelligence, specifically machine learning. The world is changing very fast and AI is being incorporated into everything. Many corporations are throwing the word “AI” into any marketing possible to impressive shareholders, this could create a potential bubble in the industry. Machine learning is a very long process of an artificial intelligence learning how to solve x. To improve machine learning you need two key ingredients 1. A large data set, 2. Feedback. Generally the model is trained on this particular data set in order to solve a particular problem. Data scientist try to feed the model as much data as possible and the more accurate the data is the better the results of the model will be. Then the model needs constant feedback from programmers that are training the model, essentially letting the model know that they are coming up with the correct results or the incorrect results, and deciding which information a model should “remember” and shouldn’t “remember”. The data set is so crucial to the model, that any poor data results in insufficient machine learning or machine learning that takes exponentially longer to eventually solve x. Data scientists constantly need new data and this is where one of the biggest problems in computer science resides right now, Finding new data. Tech companies such as Reddit and X have locked down their APIs and inflated the cost of their usage significantly since companies were essentially training on this data at a low cost for years beforehand. Sites like Reddit have massive datasets of already reviewed information by their users, which helps provide feedback to machine learning in the first place. With sites locking down their APIs so much this could possibly slow down the advancement of AI, or worse make a monopoly on just one AI model like how we see with OpenAI currently. There is an alternative way to gather data, and that is by having the models themselves generate data and learning off of this, however this can completely ruin a model as It will learn things that are completely incorrect. There are ethical problems with machine learning when training on copyrighted materials, which infringe on the original corporation's creation or even worse an individual's original work as an artist/musician. This begs the question of who does this new generative work belong to if the model was illegally trained on copyrighted material. If so it raises the questions of how we can learn from public works but “unlearn” them from the model, because to solve this dilemma you would need to create a model that is completely original and perhaps based on its own generated material. If it is possible to create a model with the capabilities of generating its own dataset that can accurately teach itself and solve its designed task, then this will be a massive breakthrough in machine learning. The future is unknown and there is much to be decided about with artificial intelligence, but I am excited as a computer scientist to see how the field progresses in the next decade.