Domain: Steam Platform

Dataset:

- https://www.kaggle.com/trolukovich/steam-games-complete-dataset
- https://www.kaggle.com/nikdavis/steam-store-raw?select=steamspy_data.csv

Relevant Information:

- Game name, release date, supported language, developer, publisher, platform, required age, genres, positive ratings, negative ratings, average playtime, median playtime, owners, price, discount price, reviews.

Learnings:

- Some of the columns' names are ambiguous and repetitive, in order to utilize the dataset we have, we need to understand what columns or rows that are significant to us and clean out the rest.
- Since these two data have some similarities and differences in their columns, we need to choose which one is more useful to us (the most recent one or the one we think is more useful to us).

Cleaning:

- There are some columns containing a bulk of information in a JSON-like form, so we need to extract the information we want by manipulating lists and strings.
- Some columns need to be renamed to a more suitable name to fit into our specialized tables.

Old Questions:

1. Platform, Requirement, Rating, Game Title: The most popular(rating, reviews) game that can play on each specific platform.
2. Genre, Rating, Release Date, Game Title: Which genre and game is the most popular in every year based on the rating.
3. Rating, Average Playtime, Median Playtime: The most-played games with worst ratings.(to be evolved).

New Questions:

1. Find the most expensive, least played (non-zero average playtime) game on each specific platform with its price and one of its supported languages is English.
2. Find the games which are the most popular. The most popular game is defined as: (1) the number of positive reviews is greater than the 80% of the total number of reviews; (2) the number of total reviews is greater than every other game on steam. There can be more than one most popular game.

3. Find all the infamous developers. An infamous developer is defined as: (1) every game they released received more negative reviews than positive reviews; (2) the number of negative reviews of each game is greater than 1000.

Schema:

- Game(Game ID, Game Title, Genre, Release Date, Developer)
- Develop(Developer, Game ID)
- Platform(Game ID, Platform)
- Rating(Game ID, Positive Review, Negative Review)
- Playtime(Game ID, Average Playtime, Median Playtime)
- Language(Game ID, Languages)
- Price(Game ID, Price, Discount Price)

Integrity Constraints:

- Develop[Game ID] ⊆ Game[Game ID]
- Develop[Developer] ⊆ Game[Developer]
- Platform[Game ID] ⊆ Develop[Game ID]
- Rating[Game ID] ⊆ Game[Game ID]
- Playtime[Game ID] ⊆ Game[Game ID]
- Language[Game ID] ⊆ Game[Game ID]
- Price[Game ID] ⊆ Game[Game ID]