1. Extract:
   1. The two datasets were from Kaggle.com and Data World about steam video games
   2. First dataset: <https://www.kaggle.com/tamber/steam-video-games/data#steam-200k.csv>
      1. This dataset is a list of user behaviors, with columns: user-id, game-title, behavior-name, value. The behaviors included are 'purchase' and 'play'. The value indicates the degree to which the behavior was performed - in the case of 'purchase' the value is always 1, and in the case of 'play' the value represents the number of hours the user has played the game. Each row contains a list of user-id, game-title, behavior-name, and value.
   3. Second dataset: <https://data.world/craigkelly/steam-game-data>
      1. This dataset is a list of several columns including QueryID, ResponseID , QueryName, ReleaseDate, RequiredAge, Metacritic, etc. The value recorded are the unique attributes of each games such as the name, release date, Metacritic rating, and price.
2. Transform
   1. First dataset: Each row contains a list of user-id, game-title, behavior-name, and value.
      1. Rename the user id, game title, behavior, and value
      2. Drop the column 0
      3. Drop all the ‘purchase’ in the behavior column to have all the hours in hours column
   2. Second dataset: Each row contains several attributes
      1. Drop most of the columns except key id, name, isfree, Metacritic, pc requirement, price, release date and DLC Count.
      2. Sort the values of Price Final in descending
   3. Merge
      1. Merge the two datasets on Name
      2. Sort the games into free and purchased games
      3. Use the price as independent variable, Metacritic rating and hours-play as dependent variable, to see if more expensive game means more people playing and better rating
         1. On average, free games have more hours played than paid games. Buying and playing more hours have a negative correlation
         2. On average, purchased games have a higher rating. Buying and higher rating have a positive correlation.
   4. Create a parent table in SQL with unique user stats
      1. Unique User with unique id for each user. Each user can have multiple games
      2. There are 41220 user id but only 9359 unique user id. On average, each user buy/play 4 – 5 games.
3. Load
   1. Load the free\_games.csv as freegames, paid\_games.csv as purchasedgames, merge\_3 as allgames, user\_id.csv as uniqueuser.
   2. User mysql to find exactly how many people play each game.