* The dataset actually includes all the games from the Steam platform at the time (initial data acquisition was in December of 2016 presumably, but list was updated a few times down the line so most likely somewhere in first quarter of 2017)
* Steam Spy database was used to acquire user data at the time, and Steam database was used for everything else. “four columns from the Steam Spy data are added by joining on the App ID” (every variable with “SteamSpy” in the name)
* “Metacritic provides a rating service for vide games, movies, television shows, and music [11]. The Metacritic score is a proprietary weighted average of various professional critics and publications, scaled from 1 to 100 [10]. It is one of the most frequently used metrics for video game quality [14].”
* Dataset includes “non-game software” which probably needs to be removed (variable called “GenreIsNonGame”)
* “the player numbers include people who were given a ‘free weekend’ to play the game [7]”
* Authors explore the Metacritic scores vs other variables, but don’t dive into differences between games that have this rating and the ones that don’t

Some caveats in data collection:

* Could not capture all data dimensions used by Steam
* “The number f owners for each game could not be acquired, but has been estimated using an outside site[7]”
* Only total number of user reviews, not distinction between “negative” and “positive” (there is a clear distinction in Steam, since users actually write them and can assign a label)
* Price recorded “as is” when data was gathered from Steam, since Steam does not store historical price data.