Overview

Steam is now the largest provider of video game digital distribution service. For any game, there is an overall review given (mostly positive, etc.) when the user clicks into the game description page. Personally speaking, I will hesitate more in buying when the review level is somewhat negative or mostly negative. With data about more than 40K steam games, developers can understand their target buyers and improve their game reviews with community-oriented adjustments. Therefore, identifying important features behind positive reviews can help developers maintain the better public image and reputation, which will bring brand effect and potential sales boost.

Research questions

The goal is to uncover the factors that lead to better review comments on Steam.

Inferential questions:

1. How does the probability of getting positive comment vary for different game features such as tag and number of achievements?

2. Does such effect differ by other features such as original price and minimum system requirement?

There are few previous researches on Kaggle on this topic or at least on steam games specifically.

Data

The data is from Kaggle*: Steam Game Complete Dataset*.

Link: https://www.kaggle.com/trolukovich/steam-games-complete-dataset

There are 20 variables and 40k+ observations. Variables include game features (tag, publisher, minimum system requirement, etc.) and sales information (original price, discount, etc.)

No need to worry about sample size and I will collapse review levels. E.g. I will collapse mostly positive and somewhat positive into a general level, positive to expediate analysis.

Project plan

Since there are multiple levels inside categorical variables, binomial multilevel logistic regression should be used.

Given four weeks left before the presentation:

Week 10: clean dataset and do EDA

Week 11: model selection and validation

Week 12: interpret and design presentation

Week 13: finish report