Steam Video Game Recommender

Springboard Capstone Project 2 Milestone Report
Greg Gibson October 2020

Problem Statement:
Based on video game votes and reviews, can a recommendation system be built to assist sales growth?

Valve Corporation, an American video game development and digital distribution company (www.valvesoftware.com), is the developer of the market dominant software distribution platform Steam (www.steampowered.com), as well as many video games.

Steam noted 90 million monthly users at the end of 2018, and Statista.com states Steam revenue in 2017 was $4.3 billion and they had 18% of the video game market. Per Steam's news update on April 7, 2020, they had nearly 1,200 new games released in 2019 that earned $10,000 in their first two weeks. Therefore, any small uptick in sales would yield great results. What if a reliable recommendation system would help gamers make confident choices on which video game would give them the most entertainment for the value?

Dataset:
Video game developers can utilize the Steam API, called Steamworks, to allow social networking and community interaction, such as friend networks, game hosting, score rankings, and posting achievements unlocked and in-game snapshots. In addition, the website collects and shares user reviews and up or down votes.

Pypi.org has a library "steamreviews" developed by Wok, https://pypi.org/project/steamreviews/#files, that handles downloading reviews by game ID. For a provided list of game IDs, each ID will be downloaded in a separate JSON file. The library accesses Steamworks user reviews and more information can be found here: https://partner.steamgames.com/doc/store/getreviews

A list of 1,000 game IDs was obtained from Steamspy, a website dedicated to estimating game sales on Steam, https://steamspy.com/api.php. The parameters through steamreviews were restricted to purchased games, English language reviews, and created within the last three years. This yielded 748 games. Each JSON file had a nested "author" column with information regarding the reviewer that was subsequently flattened into the dataframe.

The primary features to be used are the voted_up boolean and review text columns. The recommendationid is the unique field. Game ID is appid and each reviewer has a unique number in the author.steamid column. Other fields to be explored for validity and ranking support are:

- votes_up: number of users who found this review helpful
- weighted_vote_score: a rate of helpfulness
- num_reviews: number of reviews posted by this user
- playtime_at_review: user’s game playtime at time of review

<u>Exploratory Analysis</u>:

Initial counts:

- 748 game IDs
- 2,328,273 customer IDs
- 3,753,726 reviews

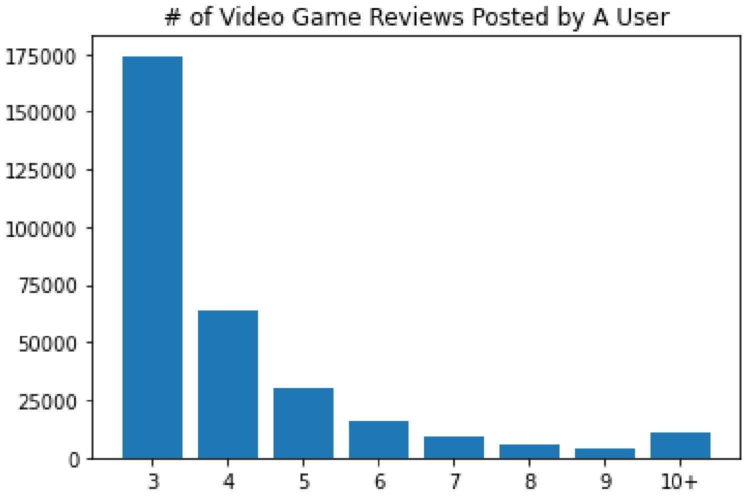
There are 6,969 blank reviews and these records are dropped. Some reviews are noted to be not in English, despite the download parameters. These rows need to be excluded as well utilizing the langdetect library.

The players:

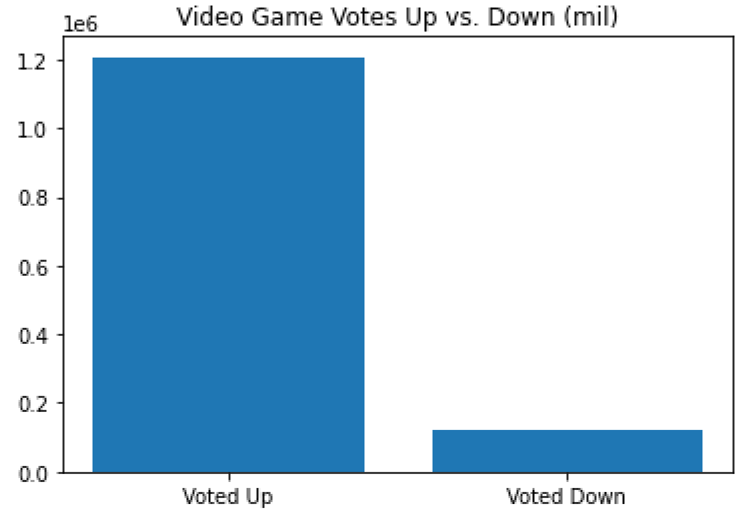
- The number of reviews per player range from 1 to 136
- There are 1,605,415 players, or 69%, with only one review and 719,150 with multiple reviews
    - There are only 75,857 players with five or more reviews which is less than 3.3% of total
    - 313,410 players posted three or more reviews, or 13.5%

This sub-group of 313K players is selected to have sufficient records for a recommender system and for go-forward analysis.

- Over half of those players, 174,185 or 55.6%, were exactly three reviews



- The players approved of video games 10x more often than disapproved



The games: