Assignment 2 Proposal
SENG5709
Daniel Koniar
4/19/20

Analysis of 2016 Steam Dataset

Introduction

For Assignment 2, I propose to analyze a steam dataset collected in 2016 by BYU University [1]. Steam is one of the largest gaming marketplaces, which offers a platform for organizing, playing, buying, and selling video games [2, 3]. With how popular Steam is across the world and my personal interest around gaming, I think it would be interesting to evaluate datasets around the popularity of certain games.

Dataset

The dataset that would be used contains 17GB of compressed data, which will need to be pruned or partitioned in order to be capable of analyzing it on my own machine. It contains games, publishers, developers, achievement statistics, and player data. With this information, I should be able perform a good variety of analytics.

Questions

Easy

1. What game has the highest average play time?
    a. What is the game publisher with the highest average play time?
    b. What is the game developer with the highest average play time?
2. What game has the highest amount of users in the dataset?
3. What user has the highest play time for a single game?

Medium

4. What user has the highest total play time?
5. What game has the highest ratio of [number of hours played] to [average hour per user]?
6. What is the average amount of users per ESRB rating?

Hard

7. What game had the biggest increase in average playtime from the earliest to latest retrieval date?
8. Did games released during the data retrieval have a higher increase in average play time? Or did pre-existing games have a higher increase in average play time?

Potential Issues

- The dataset is very large and may be difficult to perform analytics on in a timely manner. The dataset is around ~17GB compressed. Uncompressed, it is much larger (~170GB). It is not possible to load all this data into memory, which may produce thrashing.
- The dataset is from a single source, which may be hard to re-obtain data if that source goes down.

    - As of writing this document, I have had issues trying to download the dataset. This may require me to find an alternative dataset if this does not resolve itself.
- The dataset only contains a random subset of users. This means that any data done on the dataset may not be directly indicative of the total population of users. This should be added as an asterisk to the analytics.

Alternative Datasets

1. Steam Game Data From Data.World: https://data.world/craigkelly/steam-game-data
    a. This dataset would require a different set of questions

References

[1] O'Neill, M., Wu, J., Vaziripour, E., & Zappala, D. (n.d.). CONDENSING STEAM: DISTILLING THE DIVERSITY OF GAMER BEHAVIOR. Retrieved April 19, 2020, from https://steam.internet.byu.edu/

[2] Prescott, S. (2019, July 5). The most popular desktop gaming clients, ranked. Retrieved from https://www.pcgamer.com/the-most-popular-desktop-gaming-clients-ranked/

[3] Steam. (2020). Steam, The Ultimate Online Game Platform. Retrieved April 19, 2020, from https://store.steampowered.com/about/