Danielle Clarice

Steam Game Proposal

I chose Steam Store Games dataset from Kaggle. This is a great topic for both myself and the general public because there is a lot of interest in video games. There are thousands of games to choose form. It can be overwhelming and even risky to buy one when there is a chance that it wont be a good fit. Steam is the largest distribution platform when it comes to PC games. When comparing games, it is the best choice for a diverse amount of data. This data can be used to make better decision for the consumer or developer.

This dataset contains thousands of games including the price, platforms, genres, and much more. Reviews are placed in negative and positive columns; this will make it easy to compare games for the consumer. Players will find the best game to invest in based on playstyle and combability.

My friends have often downloaded a game that crash frequently due to hardware limitations. It is important to know what games are not going to crash because of RAM or graphics card.Release dates trends can be helpful for other developers to see what success other companies had at the release of their game, maybe some genres have better success during a more specific time.

Steam Store Games is interesting to me because there are a lot of games that I don't even try because I am too scared to invest in. I am always looking for a new game and it seems like a lot of work to find one. During the Covid pandemic and after becoming a parent my husband and I are not able to have much of a social life. Nighttime gaming with friends is a new activity that we are all able to enjoy. The biggest issues now are what game can we try next that is not a bust. Videogames are becoming a more popular activity, so this information can help many people.

I am using data about Steam video games. This information is found in the Steam store which I downloaded from kaggle. This type of data includes everything for the player and designer. There are columns about review, price, categories, achievements and much more.

When looking over this data my goal is to identify games that would most interest me while offering me the best bang for my buck. To do this, I will be analyzing user playtime, price and my preference in categories and genres. I believe playtime is an important indicator of player engagement and helps filter out games that are overhyped by advertising.

For the cleaning process of my data, I will begin to remove null data and standardize date format so the I can filter the release date of the games. If there an excess amount of empty or null values in a column then it will be inconsistent to use. Where there should only be numeric values, I will have to ensure they are properly formatted and only contain numeric values. This will be in columns for price, date, and playtime. Price should also be checked for outliers to catch potential input errors.

Duplicates for games need to be dropped, and string values like categories will need to be split into lists. Dropping semicolons and commas is the way that I would do this. Cleaning these columns this way will allow me to filter through categories more efficiently, and I can identify genres.

Columns that I would want to drop because they are irrelevant to my goal would be achievements, developer, and publisher. I would not need in-game achievements because that specific data is not pertinent to finding a new game that aligns with my specific genre and playstyle. While there are some

well-known developers, I would feel my results would be bias and unfair. I wouldn't want to exclude great games from unknown content creators.

I don't feel that I need to reformat my data, as I already imported it into python with panda with no issues. It is organized nicely in excel, which makes it easy to upload to my Jupyter notebook.

The data structures I will use for my project are arrays and trees. The array data structure is an easy and efficient way for me to sort and search through game data by specific values. I can use it to sort and search by specific values. I will use arrays to organize the data by playtime and price. Looking for the lowest price and highest play time will help me figure out which games offer the best value. I'll also filter the games to see how many games are under 20 dollars, then 10 dollars. On top of that I want to find games with an average 50 hours played time.

I choose trees as a data structure because they work well for organizing information in a hierarchical way. This will be helpful when I break down the genres of games. The trees parent-child relationship will allow me to narrow down genres to subgenres and then to individual games. For operations, I will build my trees by inserting nodes starting from the dataset as the root, top parent. Then the genre being the first parent tier. The spy tags being the next tier which is categories of the genre. Then I can delete nodes that are irrelevant genres and search for all nodes within a specific genre to help focus my game analysis.