Data Analysis of Steam Games

Executive Summary

Kristina Solis

2023

Career Foundry

Summary

This project consists of two datasets. The first dataset contains information about games available on Steam, a well-known video game digital distribution service and storefront. Steam was initially launched in 2003 and the data here was collected early 2023. The Steam dataset contains information including game titles, release dates, developers, genres, ratings, playtime averages and medians, and price. The second dataset contains information about game developers including company name, country, and city. The developers dataset was collected in 2019. The two datasets were merged to provide geographic data to the Steam dataset. This dataset was selected because I have long been a gamer, having grown up playing video games since the 1980s.

Data Sources

Both datasets were obtained from Kaggle. The Steam dataset was collected by Marin Bustos @fronkongames using data gathered from the Steam, SteamSpy, and Metacritic. The dataset can be found here: https://www.kaggle.com/datasets/fronkongames/steam-games-dataset. The video game developers dataset was collected by @andreshg and can be found here: https://www.kaggle.com/datasets/andreshg/videogamescompaniesregions.

Limitations and Ethics

The Steam dataset is not a full collection of all the games available on Steam. This dataset contains over 68,000 games. Because this data was collected in early 2023 (it is updated monthly) when I downloaded it, any data would not reflect 2023 trends. The collector of the game data already removed any personal identification data from the data scrape before posting it online. Developer data was limited because it did not have information for many indie game developers and there has been a boom in indie game development. Excel was used to randomly assigned developer countries for developers with this data missing. This was done to meet the requirements of this project and is not an accurate reflection of the actual location of indie game developers.

Data Cleaning and Consistency Checks

Game Developers Data (cleaned in excel)

- Deleted duplicates.
- Checked for missing data and filled in data as needed using developers' websites where possible.

Steam Game Data (cleaned in Python and excel)

- Changed column names for clarity.
- Dropped several columns that were not necessary for this analysis.
- Merged game developers data and game data together.
- Checked for mix types.
    - Present but not problematic.
- Checked for extreme values and corrected.
- Checked for missing data.

    - A lot of developer data is missing but decided to keep the rows.
- Checked for duplicates.
    - None found.

Data Profile

See data profile spreadsheet.

Column Details

See data profile spreadsheet.

Questions to explore:

- Which game developers produce games with the highest ratings?
- What genre of games have the highest and lowest number of hours played?
- Is there a correlation between game price and ratings?

Later questions:

After doing initial analysis and exploring the dataset I ended up changing focus. There are an unusually large number of Steam games with 0 hours played and the ownership level is relatively low when looking at all games.

- How many Steam games go unplayed?
- What type of games are going unplayed?
- What are the various factors that contribute to a game not ever getting played?