SteamViz - Milestone 1

SteamViz
Derin Arda Alpay, Ahmad Bilal Kakar, Thai-Nam Hoang
{derin.alpay, ahmad.kakar, thai.hoang}@epfl.ch
{300758, 296729, 369259}

I. Dataset

The dataset we are choosing is about the game distribution platform, Steam. Steam provides in-depth descriptions, reviews, video and picture examples of games, as well as other miscellaneous information including but not limited to its price, awards won, user endorsements, and such. Our dataset includes all of this information and more for over 80,000 different games. The dataset is readily available on Kaggle[1], and is ready for use. The features included within cover all our interests and more, which however causes the need to carefully trim some of them out to reduce the size of the data we are working with, since initially it is over 250 MBs. We can also observe various outliers for certain features and some features that are generally all NaNs or equivalent, which may affect the quality of the dataset. All these issues are those that should be dealt with.

In case our visualizations come to a stop due to data-related issues, mainly a lack thereof, we also are keeping in mind the possibility of inquiring a specialized website that inspects and visualizes Steam data themselves, SteamDB[2]. This website scrapes the entirety of the platform actively, and contains extensive tabular and time-series data of each game. They also mention in their FAQ that we are able to ask for their data for educational purposes. However, this data would prove far more laborious to clean and manage, and might be too expansive for this project, which is why we aren't actively pursuing it.

II. Problematic

Our main objective to visualize this data is to see which games are played the most, how well they do with respect to their price ranges, how do users react to certain types and genres of games, and to answer more questions of similar nature to then combine into a general idea of an objectively attractive and reasonable game category, and have an idea of how users react to certain features of a game. We want to proceed along the direction of how the user experience as well as the pricing of these products are affected in relation to the remainder of their features.

These visualizations will be useful for both users and developers. For the former, it will allow which games to support in early access and what to expect once they come out, or how much money they may need to put aside when certain games they are looking forward to play come out. They may also benefit by looking up how games of similar stature performed if the user reviews are not informational enough, such as for games with low number of players. For the latter, they may find some insights on how to adjust the price for their game to earn the most amount of money, as well as see current trends on gaming to decide what sort of projects they should develop next.

III. Exploratory Data Analysis

This section is covered in EDA.ipynb notebook which can be found in our GitHub[3] repository. Remark: We use an interactive plotting package, meaning if you preview the notebook on Github it won't show the figures. However, if you open the notebook on your own end, it works fine. As a precaution, you may also find the plots in the "plots" folder of our repo as pdf files, along with a plot_order.txt file telling in which order the plots appear in EDA.ipynb if needed.

IV. Related Work

As mentioned previously in the Dataset(I) section, there already is a very extensive website focused on the same topic as our project, called SteamDB[2]. Given the popularity of the Steam platform, there inevitably exists a website inspecting and providing a wide array of useful information to users, with much better data to boot, however does not explicitly cover relations between certain features overall. It instead focuses more on providing detailed information on a per-game basis. Our approach will focus on groups of games with similar features, allowing a batch-inspection of features instead of specific games. This will allow a broader view of the whole platform, which might be more useful than a case-by-case basis database for more involved individuals and groups.

However, most of our ideas mainly stem from SteamDB's website, and what we found lacking or hard-to-find within. It is organized to help gamers make the best decisions, and provides free service to all who visit. It contains more than just data as well, such as calculating the value of an account or keeping track of all announced future releases. It shows time series data updated hourly, going back for more than

10 years. Overall, it's a great website for visualizing data, and we aspire to also provide a similar service to gamers and developers alike.

REFERENCES

[1] "Steam games dataset - kaggle," https://www.kaggle.com/datasets/fronkongames/steam-games-dataset/data.

[2] "Steamdb," https://steamdb.info.

[3] "Github," https://github.com/com-480-data-visualization/project-2024-SteamViz/tree/master.