Data Visualisation COM-480 Milestone I

Xinran Tao, Ahmed Ezzo, Guanqun Liu

April 8, 2022

1 Introduction

1.1 Project Outline

Over the past few years, the gaming industry has caught up to feed the booming demand for entertainment since outdoor activities are largely unavailable during the pandemic. More interaction games and gaming platforms have brought us closer together in a time of home-staying and isolation. However, game developers focusing on technical issues may feel confused about what games are more popular in the global market and what content in-game will be more appealing to geek gamers or casual players. Grounded on this fact, our group intends to provide keen game developers with an interactive guide that targets the above perspectives. We plan to analyze and visualize library-like game datasets that document development information such as game description, genre, and developer. Besides, we will combine data with comments and scores from gamers' experiences in gaming. Game rating and recommendations will further help game newbies find their suitable pastimes. Our initial dataset exploration targets two descriptive ones: a comprehensive classification collection of published games with proper comments ("Games of All Time") and a statistical playtime and popularity dataset recording popular games on the Steam platform as of 2019 ("Steam Store Games"). In the following chapters, we will demonstrate our project aims and carry out an explanatory data analysis for the above datasets.

1.2 Project Objectives

- Conduct a comprehensive analysis of popular games from different aspects (e.g. genre, developer, interaction type, user rating, game platform).
- Sort and specialize the analysis result for technical game developers in game search or seeking development advice.
- Process the result by 3D website visualization and animation.
- Modify and optimize website contents and adjust users' visual experience in web browsing.

2 Dataset

We have chosen two datasets to explore. They can both be found on Kaggle under the title "Games of all time" and "Steam store games".

The first dataset is more general, containing information about video games in all different facets. Along with the game's name, the dataset provides the ratings, the platform, the developer, and the genre. The second dataset is more specific to games released on the Steam platform. Steam is a video game digital distribution service. This dataset contains the previous' dataset characteristics along with some others. These include the average/median playtime, users' ratings on the Steam platform, and the game's price and release date. We present a more detailed analysis of the two datasets in Section 3.

They both have a good rating on Kaggle, so we don't expect to need to do a lot of data cleaning. An initial lookup at the dataset fortifies our hypothesis. We can see that the first dataset contains 30% undefined entries. At the same time, the second one has 0 N/A entries.

Nevertheless, some data cleaning is needed to get the best possible visualizations. For instance, some game names contain Asian characters in the title and don't compile correctly in engines such as the word cloud generator. Hence, we would need to filter these carefully. Another example would be to transform the information in the dataset in a more meaningful way. When a game has multiple genres in the original datasets, they are all concatenated with a semicolon. The creator did this to avoid duplicates in the data. We can then modify the data to extract the number of times this genre appears in total (instead of having a count of the genre when it occurs with other genres).

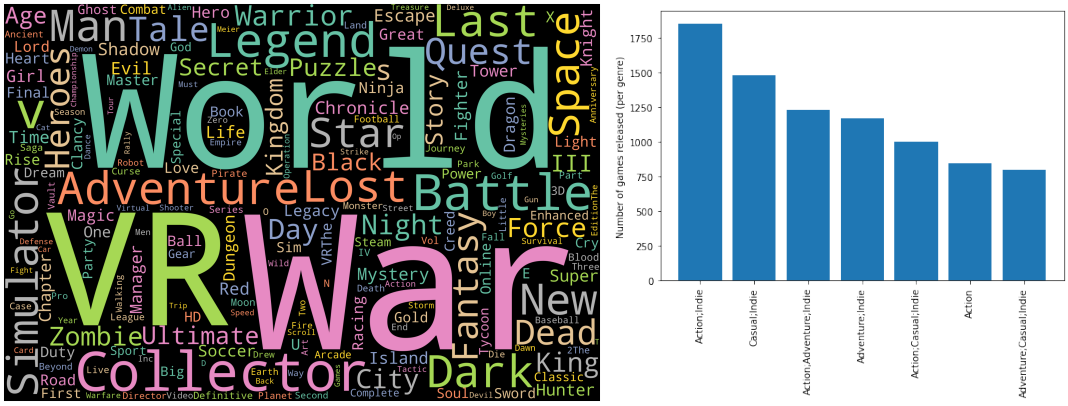

Figure 1: Visualizations of the data exploration

3 Exploratory Data Analysis

3.1 Games of All Time

This dataset contains 8831 elements, where 2898 of them are undefined entries. There is a lot of insight to investigate, thanks to this dataset. One of them is to rank developers based on different criteria. Capcom has the most released games with 150, and Nintendo comes in second with 130 games. The most beloved developer studio among the users is Vanilla Ware, with an average user score of 91.5%. We can also do all sorts of visualizations with this data. In **figure 1**, we preview a word cloud of all of the game titles in the dataset.

3.2 Steam Store Games

This dataset is very clean, with a total of 27075 elements and zero undefined entries. Instead of a general rating, the dataset provides the total number of positive ratings and the number of negative ratings of a game. More excitingly, we have access to the number of owners and the average playtime (in seconds). With this information, we can do visualizations like in **figure 1**. In this plot, we show the most popular game genres. If a game has multiple genres, we separate each genre with a semicolon. On steam, the most popular developers are Valve, with an average of 2644404 positive reviews. The second most popular developer team is PUBG Corporation, with 496184 positive reviews thanks to their big hit "PlayerUnknown's Battlegrounds". In contrast, "Choice of Games" has the most number of released games on the Steam platform (with 94 releases).

4 Related Work

The most crucial usage of this dataset is building game recommender systems. Therefore, only a few attributes of this dataset, such as *meta_score*, *user_score*, and *platform*, are listed in order to help gamers decide whether the game is worth playing. However, on top of the common numbers that regular gamers can find, a gamer developer also needs to know how certain attributes contribute to a game's popularity. In other words, what are the qualities that a good game should include.

Our approach is unique because instead of listing statistics from each attribute separately or simply picking up binary correlations between two attributes, we also aim to encode deeper connections between multiple attributes. For example, a data viewer might want to know for a specific game genre, what are the elements (e.g. art style, storytelling, game mode) that attract the user group. Our method to present this information will be creating a data point for each genre and providing filters on all other attributes. In this case, the viewer can apply different filters and see how that affects the output.

4.1 External Sources

One of our greatest inspirations comes from Unity's annual game report 2022. It has statistics based on a single attribute, such as the game genre shown in Figure 2 and it also has sections showing binary relationships between attributes, such as the one shown in Figure 3. Although Unity's report targets the marketing audience, we can still learn their way of 2D data representation to visualise simple correlations best.

In terms of presenting data with higher dimensions, we take our inspiration from a blog showing 3D data visualisation for Local Voices Network conversation data. Their website allows users to explore complicated feature correlations output from machine learning models (BERT and DeepMoji), which align with our purpose of revealing deeper connections between game data attributes.

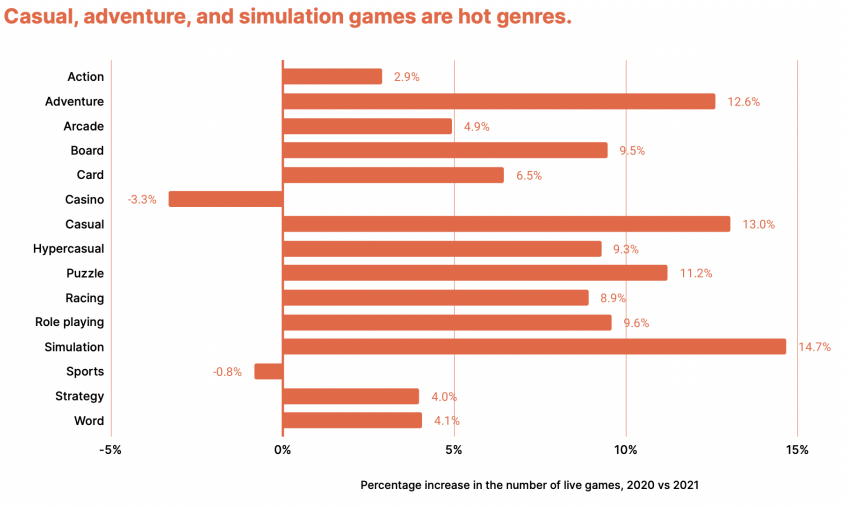

Figure 2: Hot Game Genres Chart from Unity

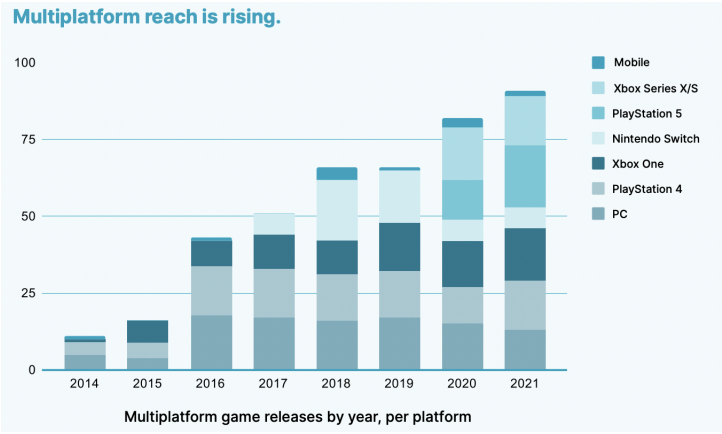

Figure 3: Multiplatform Chart from Unity