Intro of background

The booming online gaming industry is a mutually beneficial ecosystem for entrepreneurs, players and gaming developers, providing them opportunities to satiate their lucrative appetite for online gaming. In a recent report by ResearchAndMarkets, the global gaming market was valued at USD 174 billion in 2020, and it is expected to reach a value of USD 314 billion by 2026, registering a CAGR of 9.64% over 2021-2026.

Steam is the largest digital distribution platform for PC gaming serviced by Valve. It was launched as a standalone software client in September 2003 as a way for Valve to provide automatic updates for their games, and expanded to include games from third-party publishers. Steam offers digital rights management (DRM), server hosting, video streaming, and social networking services. It also provides the user with installation and automatic updating of games, and community features such as friends lists and groups, cloud storage, and in-game voice and chat functionality.

While Steam was already seeing significant growth in 2020 before COVID-19 lockdowns, video game playtime surged when people started staying home, dramatically increasing the number of customers buying and playing games. This has led to new highs for daily active users (62.6 million), peak concurrent users (24.8 million), hours of playtime (31.3 billion hours), and the number of games purchased (21.4% increase over 2019). VR Gaming on PCs has continued to show growth in 2020 with game sales up 32% year-over-year.

What topic to choose and why

As the group members all share an interest in games, we choose the games and their performances on Steam as our project topic. We want to answer the following questions:

1. What are the distributions of game genres, developers, and prices on Steam?
    a. Games are divided into many genres on Steam: action, adventure, strategy, etc. We are curious about the proportion of different genres. This may give us a hint about the trend of game market preference.
    b. Game developers are greatly diversified. By analyzing the proportion of games developed by different developers, we may learn about market concentration and get to know the key players in this industry.
    c. We are also interested in how prices are distributed on Steam. What are the higher priced types of games and why?
2. What are the most recommended games?
    a. The number of recommendations on Steam is an important indicator of trending and qualitative games. We are hoping to make use of this information to find out the games with the highest popularity.
3. What are the characteristics of those most recommended games?
    a. For those most recommended games, we are going to further analyze their characteristics and try to answer the reasons for their success.

Where and how to collect data

Finding out that the official Steam website provides an API for the users, our first plan is to use the requests methods to get data from here. However, the documentation shows that the API only provides information about news, users, and items, without any trace of games. Therefore, an alternative API with all the information we need will be used to collect data: SteamApis(https://steamapis.com/).

Firstly, we will get all games and apps in the database and extract a list of all 10210 apps' ids. Then we will use the list of id to get all softwares' details. In practice there's a problem when trying to get all ids: the request methods in python can only get the information of about 900 games, and it cannot be fixed when we change other parameters such as timeout. However, if we type the same website in Chrome, a full list of all softwares can be reached. Therefore, we will open the website in Chrome and save the contents as a json file, then open and read the file in python and extract the full list of ids.

Another problem is the limitation of the requests' frequency: for users without in-website paying, they can only request at most 100 times per minute or 5000 times per day, but we need 10000+ times of requests to get all data. In order to solve this problem, the list of ids should be divided into 3 small lists(5000+5000+210), and the frequency of the requests should be modified manually to prevent the appearance of error 429. We are still working on the data, and more details will be shown in the following data report.

What to collect and how the data should look like

For the 10,210 game data we get, each data series will look like the following example:

Appid	730
Game name	"Counter-Strike: Global Offensive"
Release date	"Aug 21, 2021"
Initial Price	14.99
Final Price	14.99
Developer	"Hidden Path Entertainment"
Category	Multi-player
Genre	Action
Recommendations	2850475