Steam Data

Chen Tianhao, Li Yufei, Xia Liang

1. Introduction

Steam is a popular game digital distribution service by Valve. In this project, we analyze a dataset about steam store which is crawled by Davis and shared on Kaggle. We downloaded the game data via https://www.kaggle.com/nikdavis/steam-store-raw . There are 3 files, 59 columns and 289MB in total. We conduct data cleaning, processing and analyzing using Pandas, Seaborn and Word Cloud library. In this project, we greatly exercised our programming ability dealing with untidy data. And we use some optimization function to make our program faster and learn more about Seaborn by setting detail parameters in our figures.

2. Cleaning the data

Considering there are too many columns in total and some information is not very effective for data analysis. We remove columns with a lot of null values, columns with unknown meanings and columns with just URLs.

```
df.drop(['controller_support', 'dlc', 'fullgame', 'header_image', 'website', 'legal_notice', 'drm_notice',
       'ext_user_account_notice', 'developer', 'publisher', 'demos', 'packages', 'package_groups', 'metacritic', 'reviews',
        'screenshots', 'movies', 'support_info', 'background', 'content_descriptors', 'content_descriptors', 'languages', 'ccu',
       'appid', 'name_y'], axis = 1, inplace = True)
```

After that we merge files by their unique steam id and remove columns which have similar information.

```
df = pd.merge(data, comment, how = 'left', left_on = 'steam_appid', right_on = 'appid')
```

3. Further Processing

For <achievements> data, we care about the total number and ignore the details. NaN values in achievements set to zero. And for the <release_date> column, we extract the key information and convert it to timestamp in pandas.

achievements	release_date
{'total': 46, 'highlighted': [{'name': 'Ace Fi...	{'coming_soon': False, 'date': '22 Jun, 2016'}

achievements
0
33

release_date
2014-10-08
2016-03-29

And we change <language_supporting> data to True-False sub-table and remerge it to the main data.

languages
English, French, Italian, German, Spanish - Spain

appid	name	is_English	is_Chinese	is_Japanese	is_Russian	is_Spanish
10	Counter-Strike	True	True	False	False	True
20	Team Fortress Classic	True	True	False	True	True
30	Day of Defeat	True	False	False	False	True
40	Deathmatch Classic	True	True	False	True	True
50	Half-Life: Opposing Force	True	False	False	False	False

We also convert data types, delete unreasonable values and rename and resort columns for better understanding.

Here are part of our main variable names in cleaning data.

Table 1. Categorical variables.

Variable	Description
appid	Unique id for each game
Name	Name of game
price	Price of game in dollar
release_date	Time stamp, YY-MM-DD
short_description	The key words in comments
developers & publishers	The company or individual develop or publish the game
achievements	Total game achievements
owners	The number of people own the game
[language]	Bool value, whether support Chinese or other languages
sup_win / mac / linux	Bool value, whether support windows, mac or linux

4. Data Analysis and Visualization

4.1 Development of Steam

We can see the development of Steam from the first figure. Few games are released before 2013 and the number of games has exploded after 2017. Because our data is to July 2019, so the line falls at the end.

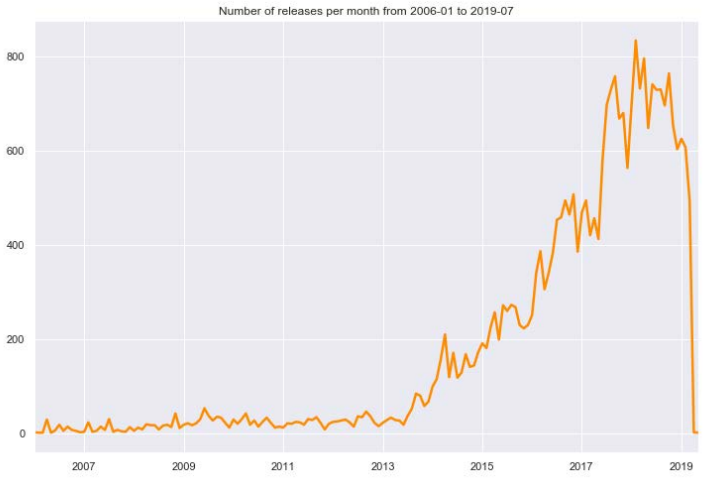

Figure 1. Development of Steam, sum of game released by time series

4.2 Popular games released by years

First we define a ‘popular’ game has above twenty thousands owners. And that is the orange part in the figure.

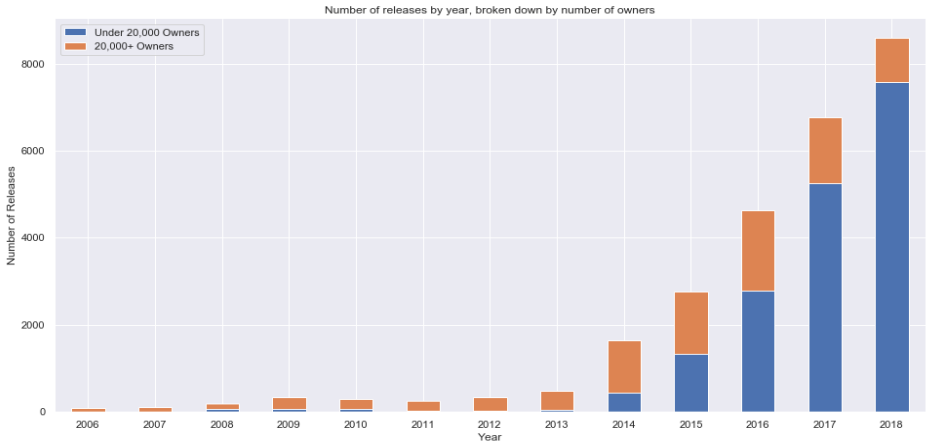

Figure 2. Popular games released by years

4.3 Top 5 popular games

name	owners	genre	developer	publisher	price
PLAYERUNKNOWN'S BATTLEGROUNDS	75000000.0	Action, Adventure, Massively Multiplayer	PUBG Corporation	PUBG Corporation	29.99
Counter-Strike	15000000.0	Action	Valve	Valve	9.99
Counter-Strike: Source	15000000.0	Action	Valve	Valve	9.99
Grand Theft Auto V	15000000.0	Action, Adventure	Rockstar North	Rockstar Games	29.99
The Elder Scrolls V: Skyrim	15000000.0	RPG	Bethesda Game Studios	Bethesda Softworks	19.99

Figure 3. Top 5 popular paid games

name	owners	genre	developer	publisher	price
Dota 2	150000000.0	Action, Free to Play, Strategy	Valve	Valve	0.0
Counter-Strike: Global Offensive	75000000.0	Action, Free to Play	Valve;Hidden Path Entertainment	Valve	0.0
Team Fortress 2	35000000.0	Action, Free to Play	Valve	Valve	0.0
Warframe	35000000.0	Action, Free to Play	Digital Extremes	Digital Extremes	0.0
Unturned	35000000.0	Action, Adventure, Casual, Free to Play, Indie	Smartly Dressed Games	Smartly Dressed Games	0.0

Figure 4. Top 5 popular free games

4.4 Supporting System

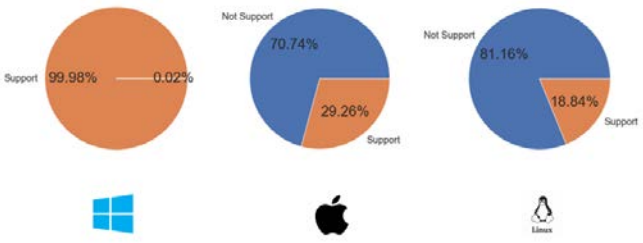

Figure 5. Supporting System Porportation

4.5 Percentage of Chinese supporting

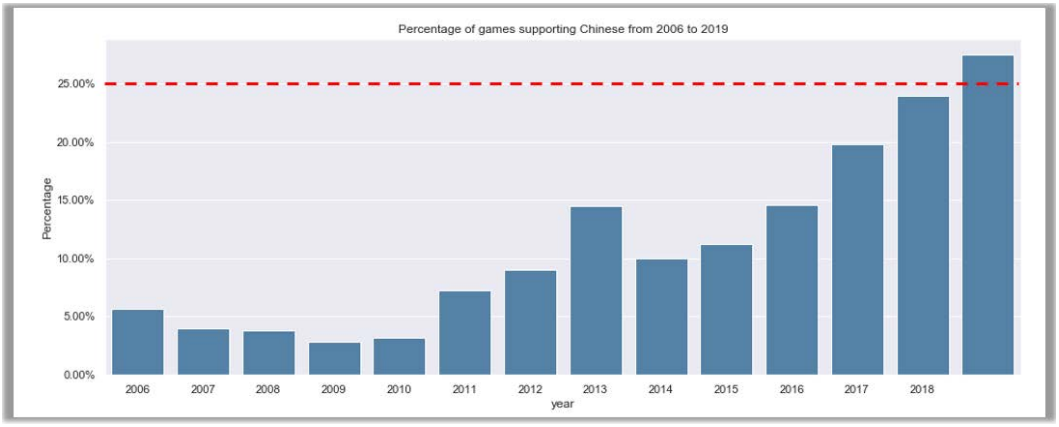

Figure 6. Percentage of Chinese supporting

4.7 Word Cloud in description

4. Conclusion

By employing python, we conduct data exploration on the Steam Data. Data analysis reveals Steam store is a rapid development platform in recent years. We also discuss about the properties of popular game and give how percentage of Chinese supporting change by year. Due to the limitation of the data set content, we did not focus on prediction or other fancy tools.