**Steam Data**

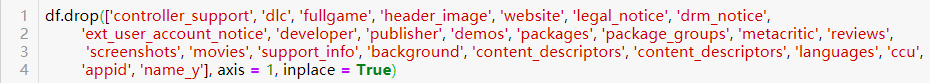
**Chen Tianhao, Li Yufei, Xia Liang**

1. **Introduction**

Steam is a popular game [digital distribution](https://en.wikipedia.org/wiki/Digital_distribution) service by [Valve](https://en.wikipedia.org/wiki/Valve_Corporation). 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.

1. **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.

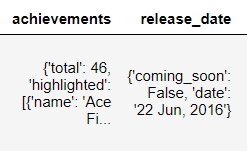
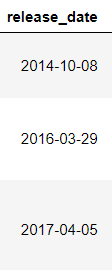


After that we merge files by their unique steam id and remove columns which have similar information.

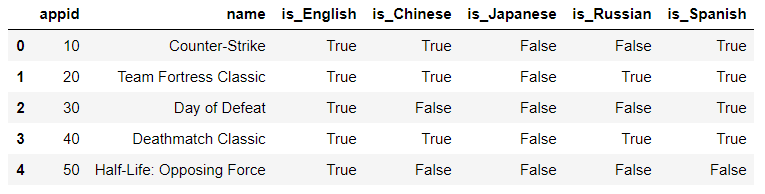
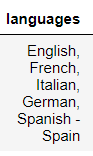


1. **Further Processing**

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.

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



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.

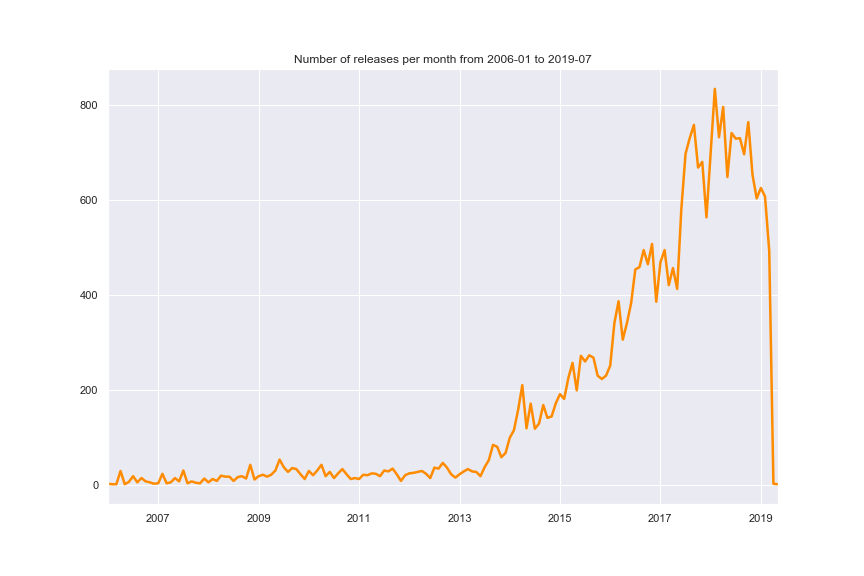
**Table 1.** Categorical variables.

|  |  |
| --- | --- |
| 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

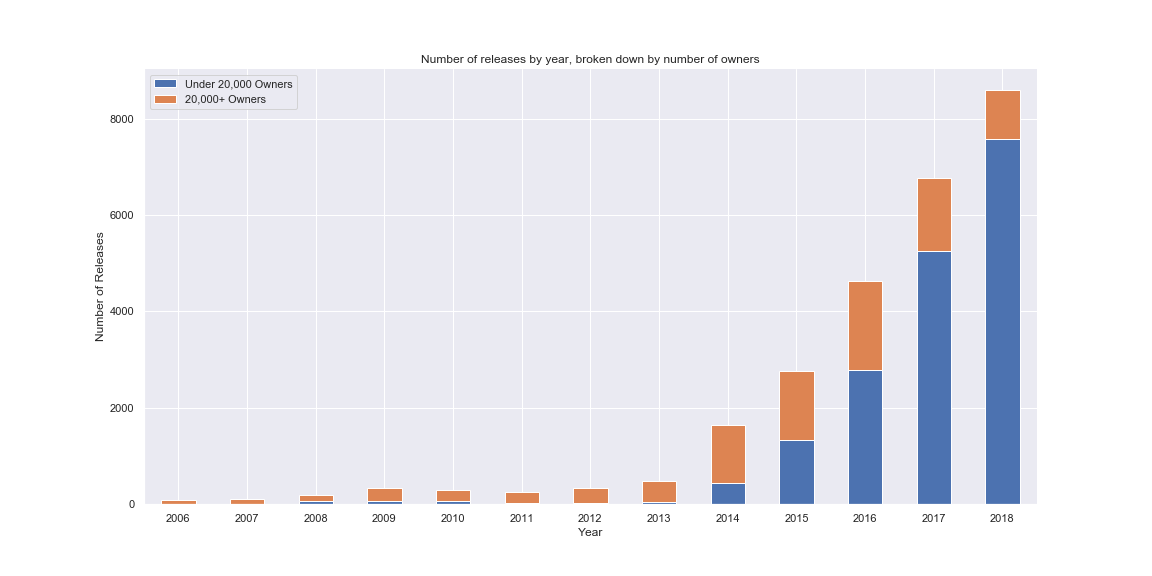
We can see the development of Steam from the first figure. Few games are released before 2013 and the number of games has exploded after 2017. Because our data is to July 2019, so the line falls at the end.



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

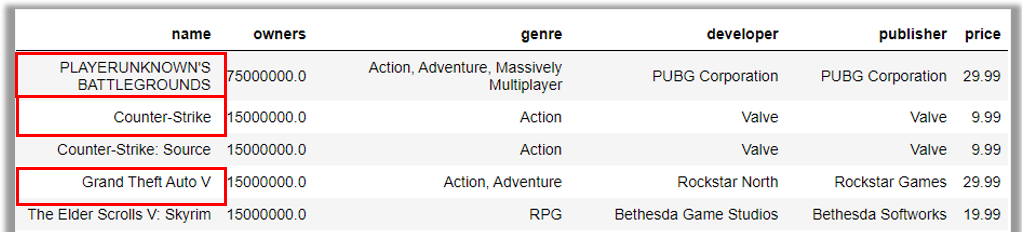
4.2 Popular games released by years

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

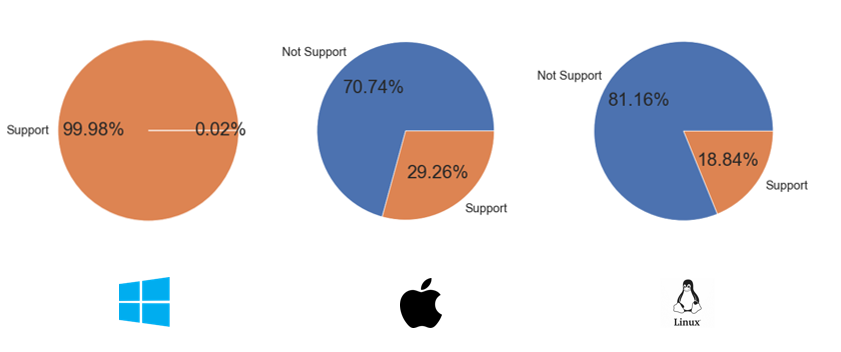


**Figure 3.** Top 5 popular paid games



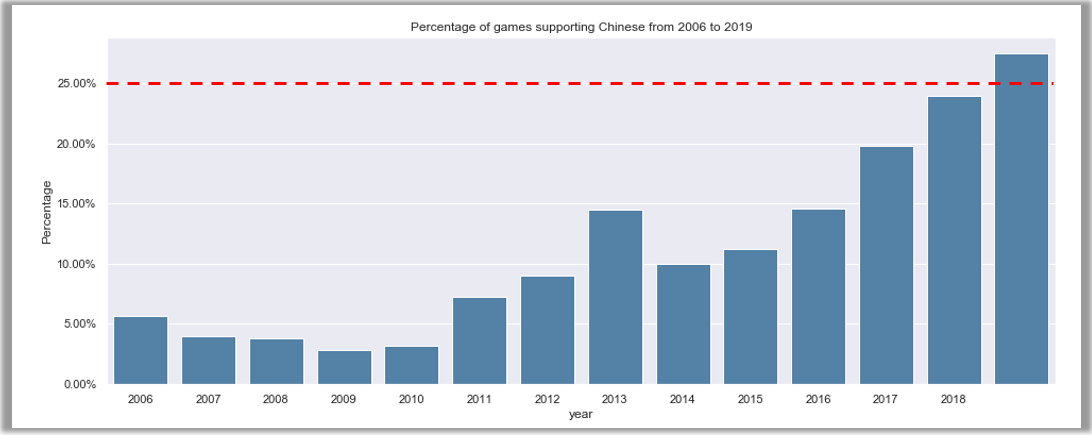
**Figure 4.** Top 5 popular free games

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



1. **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.