SI 618 Project 2 Report

Exploratory Data Analysis of video games on Steam

Sijun Tao (sijuntao@umich.edu)

1 Motivation and Summary

Currently, the video game industry is in a period of rapid growth. With billions of dollars of revenues and a large number of players worldwide, the video game platforms expand rapidly, among which Steam is one of the most famous. Steam is one of the most popular platforms on which people can search and find the games they are interested.

As one of the Steam users and someone interested in being a game developer, I'm focusing my project on evaluating the games on Steam. What type of games are most popular? Which developer has the most games with high ratings? What game price setting is common among the games?... To solve these, this project will perform time series analysis, exploratory analysis and clustering analysis to investigate the potential relationship between the game factors such as price, genre and ratings.

In details, I will focus on three main problems:

1. How does time affect the game price setting and user's activities?
   - Does there exist any tendency of the game price setting?
   - What is the average user play time over the months?
2. Do game prices and ratings have an impact on market performance?
   - How much is the game price commonly set?
   - What is the game price for the top 10 developers who have most developed games?
   - Is there a correlation between positive ratings ratio and the game owner number?
   - Does higher positive ratings ratio indicate higher user play time?
3. Do things that have similar game ratings share any characteristics?
   - Are there certain genres of games that have similar ratings? Is there any pattern?
   - Are there certain developers among the top 16 developers who developed most games with similar ratings?

2 Data Source

The dataset I used for this project is from Kaggle and is in CSV format. The link to the dataset is https://www.kaggle.com/datasets/nikdavis/steam-store-games. This dataset contains games released prior to May 2019 on Steam. It has 18 columns, including information about the video game itself and the users' activity on it such as ratings and play time. The data includes games released from 1997 to 2019, and it has 27075 records. Some important fields I used are:

- **Release date:** It is the timestamp that indicates the release date of the game
- **Developer:** The developer of the game
- **Genres:** The labels that the game with
- **Positive ratings:** The number of positive ratings of the game

- **Negative ratings:** The number of negative ratings of the game
- **Average playtime:** The average playtime of the users playing the game
- **Owners:** The number of users who purchased the game
- **Price:** The price of the game

3 Methods

In this section, I will describe the manipulation and preprocessing I did to the data.

3.1 Question 1: How does time affect the game price setting and user's activities?

- **Manipulation**:

  To do the time analysis, I first transformed the "release_date" attribute to be in timestamp format. Then, I add "year", "month", "day" attributes using the data in "release_date" attribute so that I can visualize the data on different time scale. Since the records are on a daily scale, I averaged the prices and median user play time of the games released in that month to get the game prices and median user play time on a monthly basis. Similar for game prices on a yearly basis.

  According to my observation, the yearly basis plot cannot reveal the detailed vibrations over the time. Therefore, I choose the data on the monthly basis to visualize.

- **Abnormal data processing**:

  There are several records with the "median_ playtime" attribute greater than 1500. Since such cases only happens to 2 records among the 27075 records, I filter out these 2 records to make the trend in the plot clearer.

- **Challenges**:

  Although I filter out some outliers with extreme values, the time effect is not clear enough during some period such as 1997-2002 since "median_playtime" and "price" attributes have a dramatic rise or fall. This may be because the volume of data in such periods is not large enough. Due to the relative lack of the data in certain periods, the changes of the variables cannot be interpreted reasonably.

3.2 Question 2: Do game prices and ratings have an impact on market performance?

- **Manipulation**:

  For this question, I did several data manipulation separately for its sub-questions. For the distribution of game price, I filtered the data to get the records for the games of the top five developers with the largest number of developed games. For the exploratory questions about the relationship between the variables, since the value of "owners" attribute is presented only as a range, such as 0-20000, I set a random integer within this range as the number of estimated owners of the game for further visualization.

- **Abnormal data processing**:

  Since I want to find the common pattern for the influence of game ratings on market performance, I filtered out records with extreme values of "estimated_owner" and "average_playtime" attributes. In this way, the outliers cannot have a large effect on the fitting and the calculation of correlation.

- **Challenges**:

Similar as question1, the data volume in different periods varies greatly. Due to this unevenness, the weights of data in different periods for the final fitting also vary greately so that the fitting result is not accurate enough. To reduce the impact, I chose to fit the data from 2005-2019 where the data volume is relatively uniform to eliminate the time effect in this analysis.

3.3 Question 3: Do things that have similar game ratings share any characteristics?

- **Manipulation**:
  To do the clustering, I needed to re-indexed the data. Since one game have several genres, I first explode the data so that each row has only one genre. Then, I grouped the rows with the same genre together and calculate the average positive ratings and negative ratings. Finally, I set the "genre" attribute as the index and use the average positive and negative ratings data for further clustering. For the clustering of developers, I used similar steps to get the re-indexed data for further clustering.
- **Abnormal data processing**:
  I dropped the records with nan value of positive ratings and negative ratings. I also filtered out the rows with extreme values.
- **Challenges**:
  The clustering result for the genres is relatively sparse. The direct clustering for all the genres may not be appropriate enough. More refinement groups are needed before the clustering to get the better performance.

4 Analysis and Results

In this section, I will describe the code workflow and present visualizations and corresponding analysis of the results. The visualizations are in the file *sijuntao_si618_project2.ipynb*.

4.1 Question 1: How does time affect the game price setting and user's activities?

Workflow: First, I did the manipulation to the data as described in section 3.1. Then, I used ggplots with year intervals to perform the effect of time on the variables "price" and "median_playtime".

- **Does there exist any tendency of the game price setting?**

  According to the figure 1, we can find that the change of game price can be divided into two stages. The first stage is 1997-2005 where the game price changes dramatically. It is obvious that the number of data points in this period is relatively low compared to that from 2006-2019, which means that the data from 1997-2005 is not very informative to be convincing enough. This may be because the number of developed games is not large, or many games released in this period are not available on Steam.

  For the second stage, 2006-2019, the number of data points is large enough for us to analyze. The game price has an overall tendency of increasing during 2006-2013, decreasing during 2013-2018 and increasing during 2018-2019. Besides, during 2006-2012, the amplitude of the vibration of the game prices is greater than that after 2012.

This may be because the game market is relatively stable after 2012. After 2012, the game prices in October, November and December are usually larger than the prices in other months of a year.

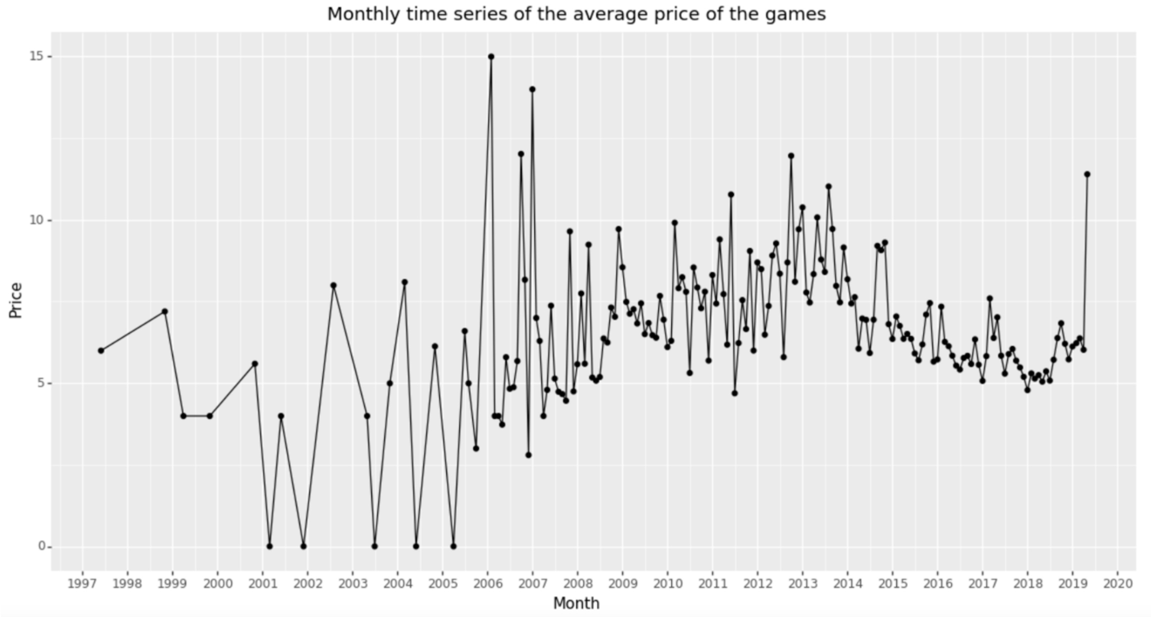

Figure 1: Monthly time series of the average price of the games.

- **What is the average user play time over the months?**

Through the figure 2, we can find that the feature "user playtime" behaves similarly as the game price. The number of data points from 1997-2006 is not large enough to provide

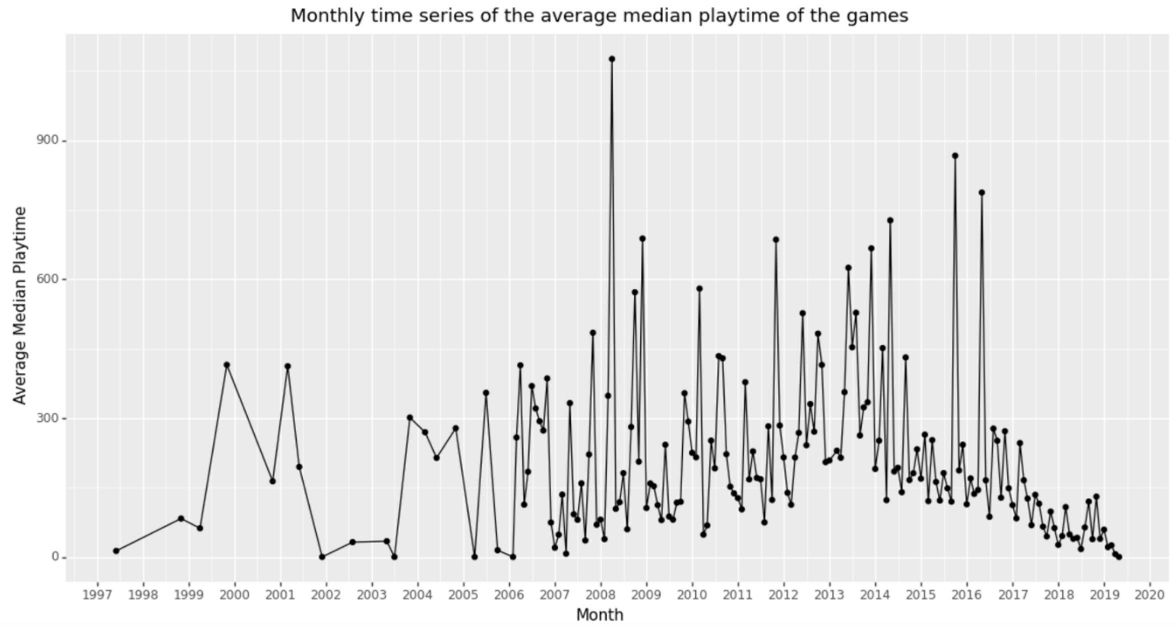

Figure 2: Monthly time series of the average playtime of the games.

convincing analysis. However, we can find that the user playtime has an overall increasing tendency before 2014, which may be because the development and popularization of the Internet. After 2014, the user play time has an overall steady decline except two months between 2015 and 2017. This may be because the number of game platforms becomes larger and user tends to play the game on other platforms.

4.2 Question 2: Do game prices and ratings have an impact on market performance?

Workflow: First, I did the manipulation to the data as described in section 3.2. Then, I used distribution, bar plot and reg plots to explore the relationship between the different factors and the market performance such as game owner number and user play time.

- **How much is the game price commonly set?**

The statistics about the game prices is given in Table 1 as below:

Indicator	Game Price
mean	6.078193
std	7.874922
min	0.000000
median	3.990000
max	421.99000

Table 1: The statistics of the game price.

From table 1, we can find the game prices varies from $0 to $421.99. According the mean and median of the price, the price of most of the games concentrates around $4, which is much smaller than the largest game price.

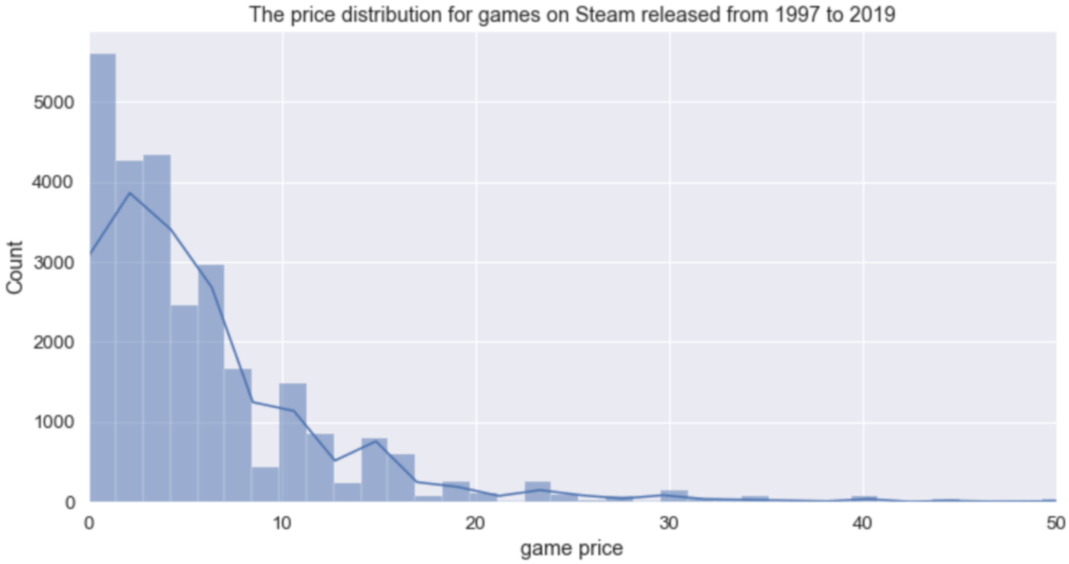

Figure 3: The price distribution.

The price distribution confirms the statistics obtained before. Most of the games set the price to be relatively low, around $3-$4.

- **What is the game price for the top 10 developers?**

Since I want to get the game price settings of the experienced developers, that is, the developers who have the largest number of developed games, I plotted the distribution of their price settings as shown in figure 4. It turns out that eight out of the ten top developers set the game price lower than 5, which is consistent with our previous finding. However, Ripknot Systems and KOEI TECMO GAMES CO., LTD have relatively high game prices. Besides, the latter one also has a broad range of game prices varies from $14 to $50, which indicates that the aim of this developer is the high-end game market.

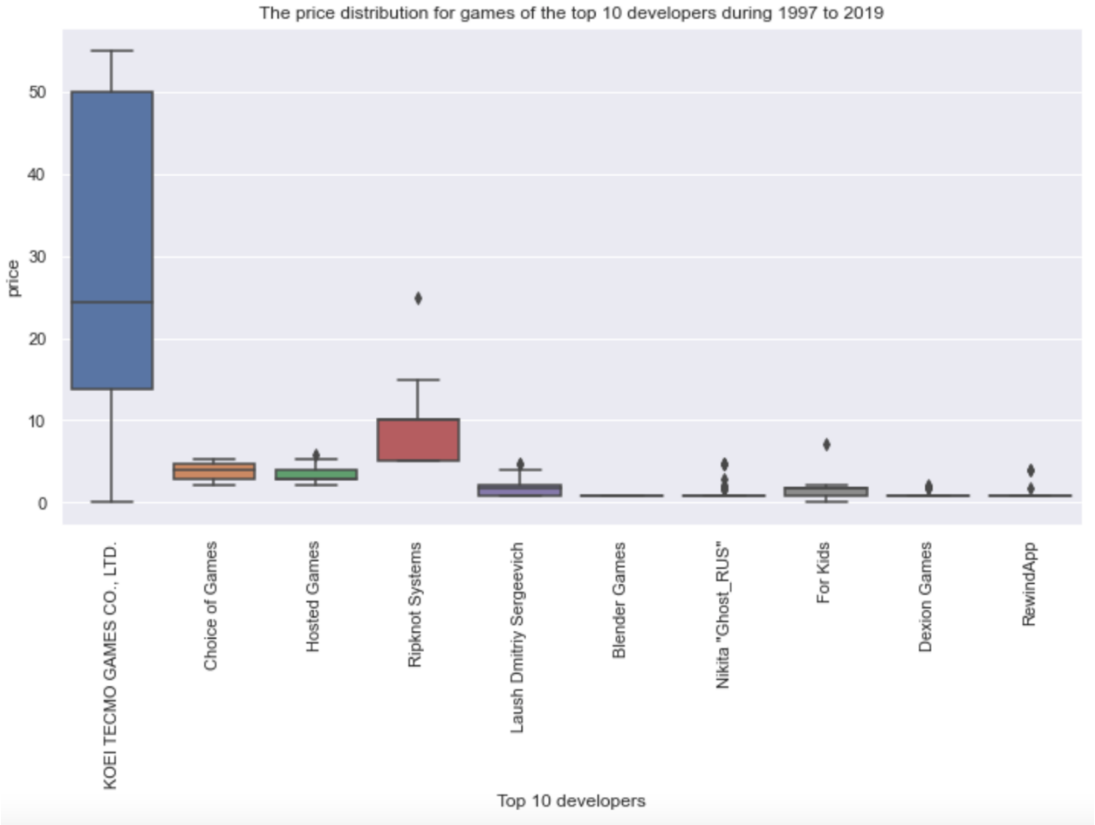

Figure 4: The price distribution of games of the top 10 developers.

- **Is there a correlation between positive ratings ratio and the owner number?**

As shown in figure 5, the linear fitting line given by reg plot indicates that there exists a positive relationship between positive ratings ratio of the game and the owner number. In addition, we can also see that most of the games with more game users, that is, games with more than 20 million users, have a positive rating ratio greater than 0.6. The higher the positive ratings ratio the game have, the larger the number of the estimated owner is.

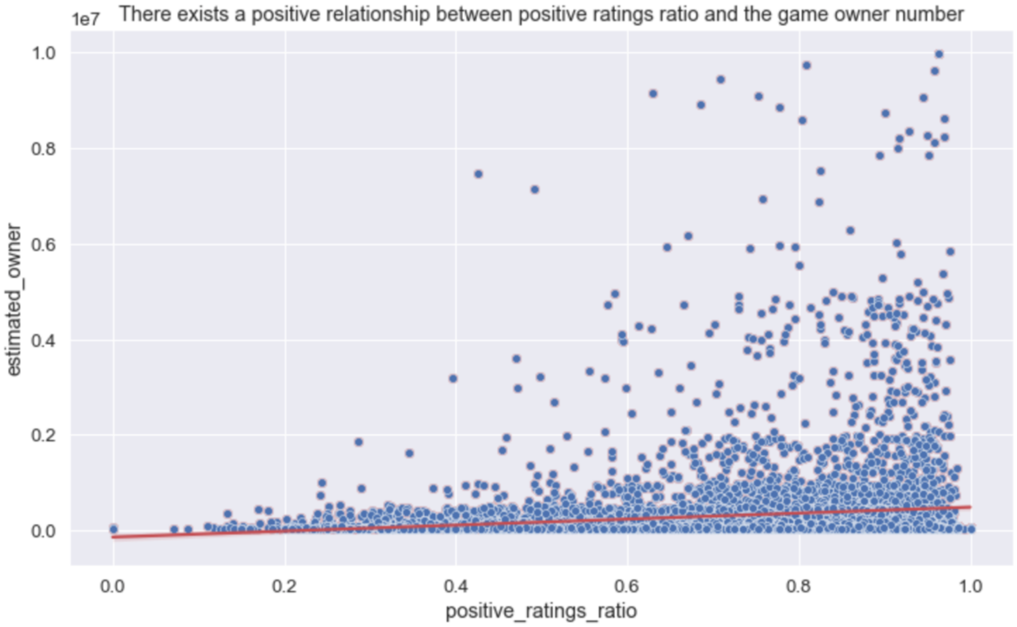

Figure 5: Positive ratings ratio vs. owner number.

- **Does higher positive ratings ratio indicate higher user play time?**

As shown in figure 6, the linear fitting line given by reg plot indicates that there exists a positive relationship between positive ratings ratio of the game and the average user playtime. It is obvious that the data points are concentrate around the right side of the plot and the data points with relatively high average playtime are more likely to appear on the

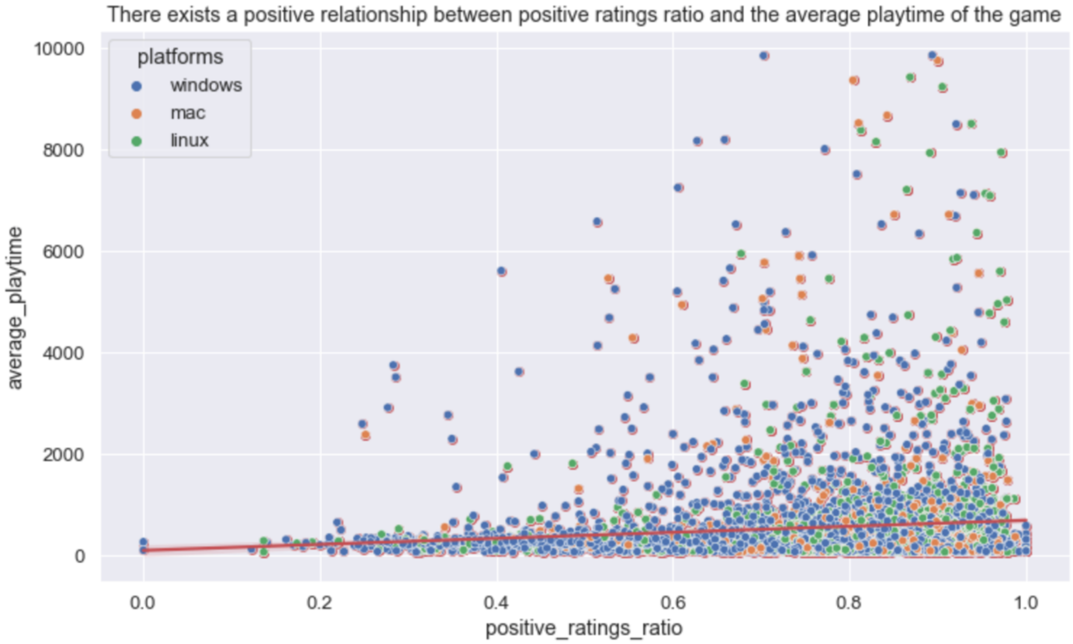

Figure 6: Positive ratings ratio vs. average playtime.

right side of the plot. Therefore, we can infer that a higher rate of positive ratings means a higher amount of time spent playing.

Besides, from the color of the data points, we can find that the platforms for most games is windows. The type of the platforms does not have much influence on the relationship between positive ratings ratio and user playtime since the distribution of the points of different colors are similar.

4.3 Question 3: Do things that have similar game ratings share any characteristics?

Workflow: First, I did the manipulation to the data as described in section 3.3. Then, I used dendrograms to cluster the genres and developers according to their similarities on the user ratings.

- **Are there certain genres that have similar ratings? Is there any pattern?**

According to figure 7, we can find that the clustering of the genres can mainly be divided into three groups. The first orange group has only two genres, "Free to play" and "Massively Multiplayer". This indicates that the players are more willing to play the free games with their friends. The second green group has several subgroups. The genres in these subgroups are generally all about professional skills learning, the games with these labels are more likely to be educational games. The third red group also has several subgroups. However, the genres in the third red group are about the classification of the entertainment games. From the clustering in the third group, we can find that the "Action" and "RPG" share more similarities. This maybe because RPG type action game is one of the most popular games in the market so the two genres usually appear at the same time.

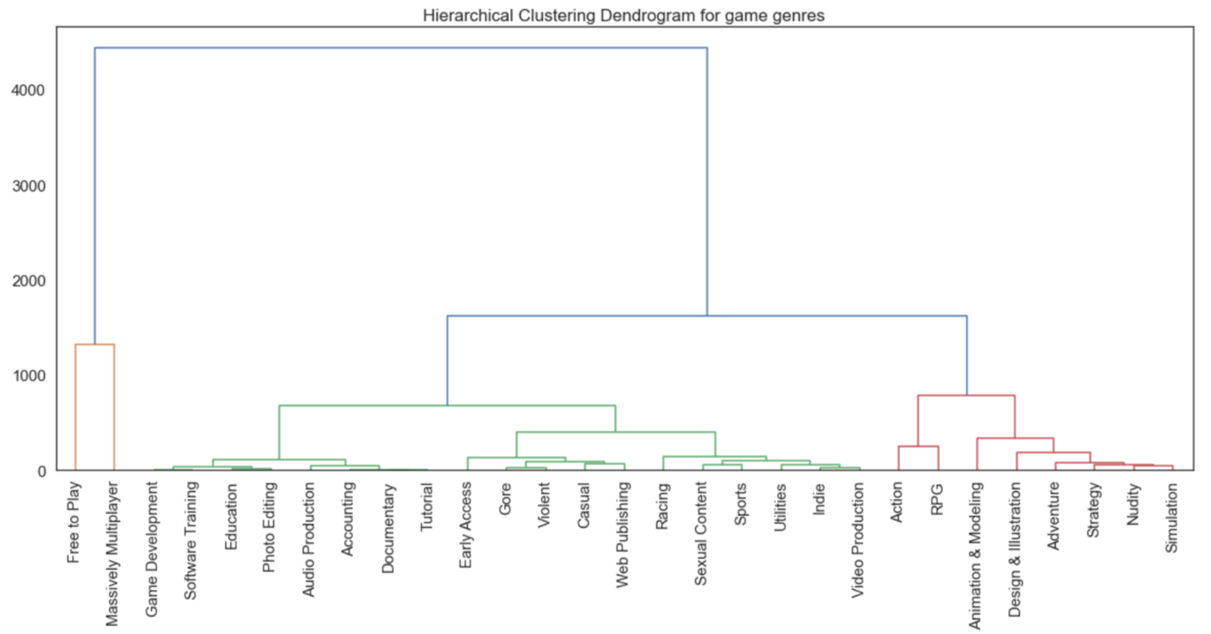

Figure 7: Hierarchical clustering dendrogram for game genres.

- **Are there certain developers among the top 16 developers who developed most games with similar ratings?**

According to figure 8, we can find that the genres are divided into three groups. The first orange group has two subgroups. Most of the developers in this group only develop games with the same type. For example, most of the games developed by RewindApp are racing games, while most of the games developed by Hosted Games are 3D adventure puzzle games. For the second green group, Ripknot Systems and Warfare Studios share more characteristics in terms of game ratings. For the third red group, the number of developers in this group is relatively low. The similarities between the developers are relatively low so the developers may not share much characteristics.

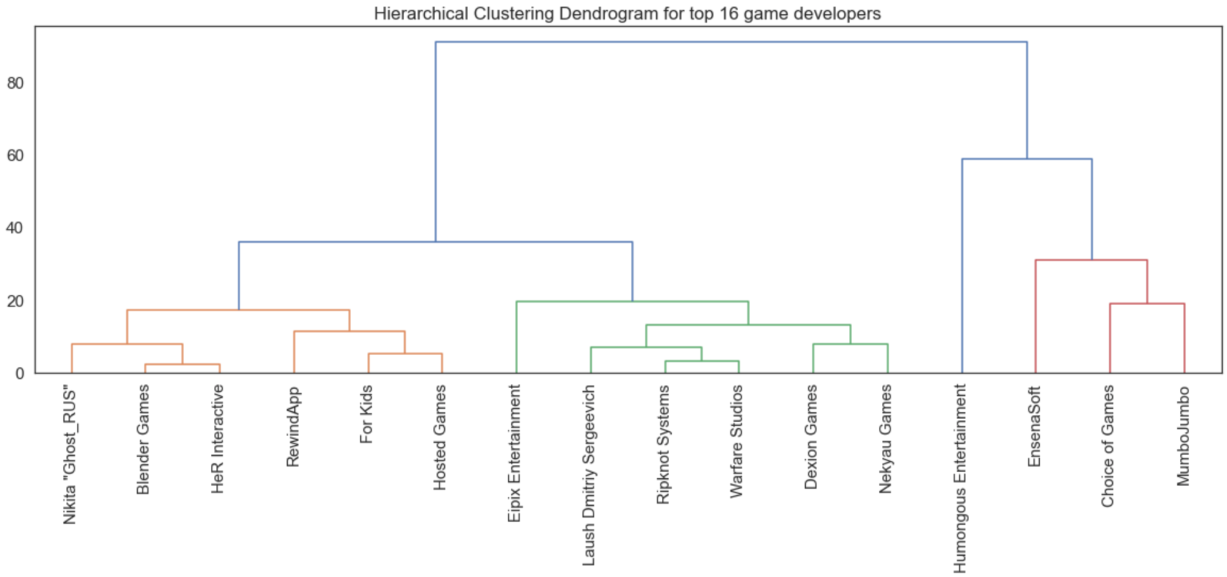

Figure 8: Hierarchical clustering dendrogram for top 16 game developers.

5 Conclusion and Discussion

Overall, this project gives me an insight of the data exploratory analysis in real practice. From the visualization and analysis above, I learned about the game related data and the relationship between the features, which may help my future career as a game developer. Nevertheless, I think there exists room for improvement in this project.

First, the original volume data is uneven. The volume of data is very small during 1997-2005 compared to that from 2006-2019. This leads to the fact that the data after 2006 has more weight in fitting, resulting the result to be not accurate enough. Second, some visualization methods of analysis are crude. Further classification and refinement of the data can lead to more detailed conclusions. If I had more time, I would investigate oversampling techniques to fit various models and do further classification and refinement of the data to get better analysis.