ISSN 2959-6157

Visualization and Analysis of Consumer Data in Gaming

Yixuan Du

School of senior, Nanjing Dongshan Foreign Language School, Nanjing, China
*1807040233@stu.hrbust.edu.cn

Abstract:

Games represent a structured form of play, typically engaged in for amusement, pleasure, or educational purposes. These activities span a diverse range, including board games, card games, video games, and outdoor physical games. Central to their design are defined rules, objectives, competitive elements, and interaction among participants. Universally cherished across all ages and cultures, games serve not only as entertainment but also as dynamic tools for interaction and learning. Each game consists of numerous components, and the alignment and execution of these elements can determine whether a game is perceived as successful or not. Crucially, consumers act as the primary indicator of a game's reception. By analyzing game-related data, one can gain insights into consumer preferences and behaviors. This involves collecting extensive datasets, performing rigorous analysis, and employing visualization techniques to elucidate patterns and trends. Additionally, various models are applied to discern outcomes, enhancing understanding of what aspects resonate with users and drive engagement. This comprehensive approach not only aids in optimizing game design but also in tailoring experiences to meet evolving consumer demands.

Keywords: Computational modeling; Games; Performance evaluation; Consumer preference.

1. Introduction

Video games have been a vibrant part of global culture for over 70 years, evolving significantly since their inception. Throughout this period, developers have continually leveraged their creativity, progressively crafting games that are not only increasingly complex and diverse but also more engaging. In 1972, a milestone was reached when the first game console was installed in a bar, operable with a coin [1]. This machine met with overwhelming success, frequently malfunctioning due to an excess of coins inserted by enthusiastic players. This incident marked the beginning of what many consider the golden age of gaming, a period when the industry's potential for profit became widely recognized. Today, the video game industry offers thousands of games across various genres on multiple platforms. These games vary greatly in popularity, visibility, and reception. They present distinct themes, styles, and core concepts, each tailored to different audiences [2]. This project aims to explore the multifaceted factors influencing game popularity. It will examine how the nature of the game, the platforms they are released on, and the timing of these releases affect their success. By analyzing trends across successful and less popular games, this study seeks to uncover patterns and deviations that could inform future game development and marketing strategies. Further, this research intends to employ advanced analytical techniques to dissect the complex interactions between game characteristics and consumer preferences. Through detailed data visualization and model implementation, insights will be derived to enhance understanding of what drives engagement and sustains interest in this dynamic field. This comprehensive approach will not only help in pinpointing the elements that contribute to a game's success but also assist developers in designing experiences that resonate with diverse gaming communities worldwide.

1.1 Data Collection

The data we collected was a dataset containing a list of video games with sales greater than 100,000 copies. It was generated by a scrape of vgchartz.com. It is from the file backloggd_ games. csv which was downloaded from the website Kaggle [3]. The tabulation contained 60,000 game samples and each sample has 13 columns of parameters.

However, we could not begin the data analysis yet. Most of the parameters in the file could not be understood by the computer, and will cause error when we tried to plot a graph with them. For example, "2023-May-12" is a parameter of the date when the game was released. A person can easily understand that the game is released in 12th of May in 2023. But a computer does not know what 'May' mean, and is not able to arrange the dates from past to present, not without our programming. Apart from

releasing dates, the genres, platforms and companies of the games are also not able to be used for graph plotting directly.

1.2 Data Processing

In order to make the data comprehensible for computers, we decided to process it with Python, a powerful programming language.

To render the data comprehensible for computational analysis, we employed Python, renowned for its robust capabilities in data processing. The treatment of data related to game genres and platforms follows a similar methodology, necessitated by the combined nature of these variables [4]. For instance, the genre data for the game 'Elden Ring' is presented as ['Adventure', 'RPG'], which aggregates the genres into a single list. However, computers require these genres to be distinctly recognized, such as ['Adventure'] separately from ['RPG']. To address this, we developed a program that iteratively processes each game's genre data. This program executes a loop 60,000 times, where in each iteration, it extracts the genre data, splits it into individual, comprehensible components, and subsequently stores these components into an initially empty list. Importantly, the program includes a verification step before adding each new item to the list to prevent the inclusion of duplicate entries [5]. The specific code used in this process is illustrated in Table 1, showcasing the approach taken to ensure that each genre is uniquely identified and accurately cataloged for further analysis.:

Table 1. The specific coding

```
#find out all the genres appeared in df_game(some games have several genres)
Genres = []
repetition = False
for i in range ( 0, 60000 ) :
    CurrentGenres = df_game.iloc[ i, 6 ]
    CurrentGenres = CurrentGenres[2 : ( len ( CurrentGenres ) - 2 ) ]
    CurrentGenres = CurrentGenres.split( "', '" )
    if CurrentGenres != ['''] :
        for j in range ( 0, len ( CurrentGenres ) ) :
            repetition = False
            for k in range ( 0 , len ( Genres ) ) :
                if CurrentGenres[ j ] == Genres[ k ]:
                    repetition = True
            if repetition == False:
                Genres.append ( CurrentGenres[ j ] )
```

While dealing the data of releasing date, our original plan was transfer the date into a float. Like "2023-May-12" will become "2023.05" after processing [6]. Unfortunately, python crashed during the process because of the limitation of our computer and the giant quantity of data. Our plan B was to only keep the year of releasing date and it worked. The specific coding in Table 2:

Table 2. The specific coding

```
df_graph3 = pd.DataFrame ( np.zeros ( ( 100, 23 ) ) , columns = Genres )
df_date = pd.DataFrame ( np.zeros ( ( 100, 1 ) ) , columns = [ 'Date' ] )
df_graph3 = pd.concat ( [df_graph3, df_date], axis = 1 )
date = 0.0
count = 0
for i in range ( 0, 60000 ) :
    #for j in range ( 0, 12 ) :
        if ( '2' in df_game.iloc[ i, 2 ] ) or ( '1' in df_game.iloc[ i, 2 ] ) :
            date = int ( df_game.iloc[ i, 2 ][ 8 : 12 ] )
            repetition = False
            for k in range ( 0, count ) :
                if date == df_graph3.iloc[ k, 23 ] :
                    repetition = True
            if repetition == False :
                df_graph3.iloc[ count, 23 ] = date
                count = count + 1
```

2. Data analysis

2.1 Popularity of games of different genres

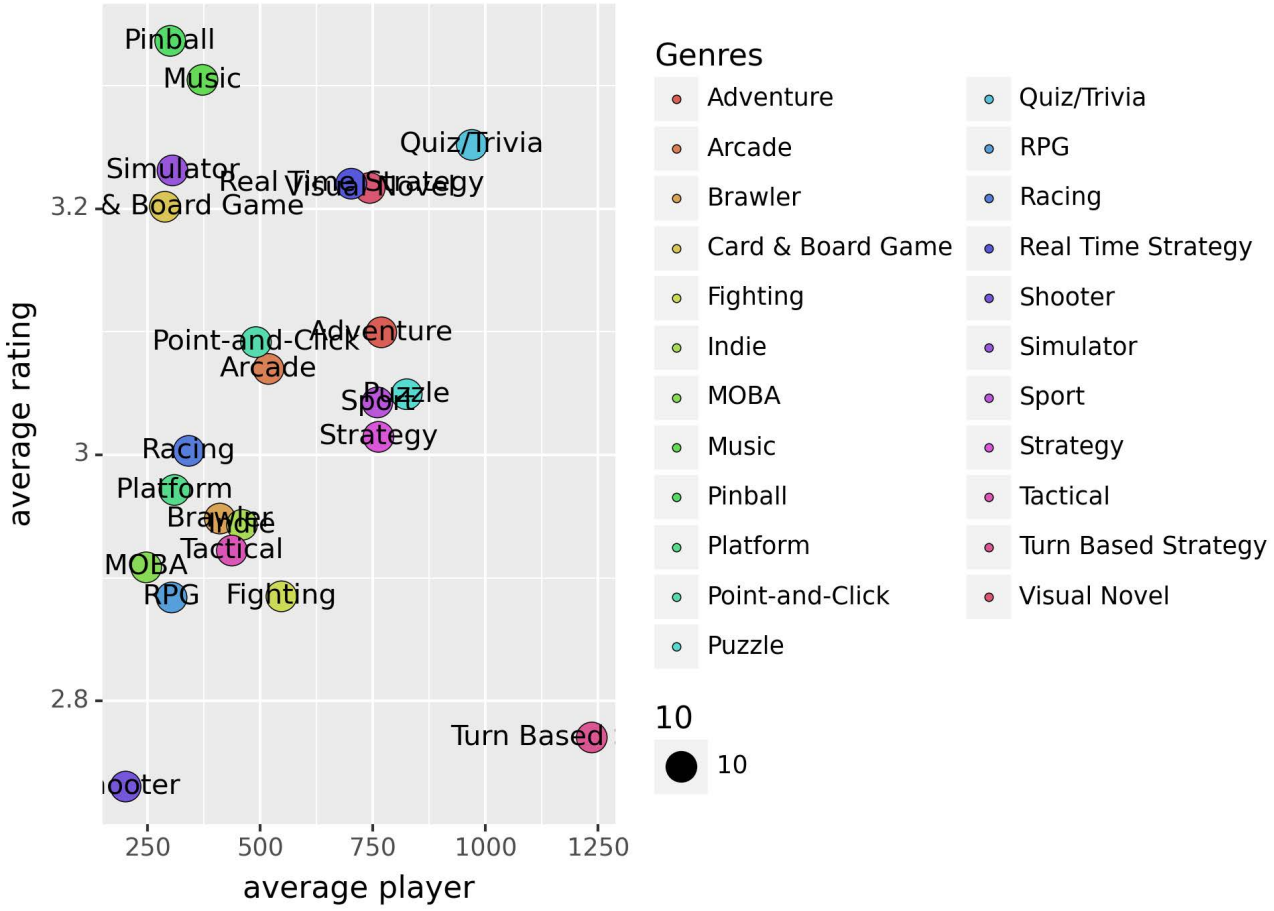

Fig. 1 Popularity of games of different genres (Photo/Picture credit: Original).

The scatter diagram above shows the relationship between average rating and average player count of different genres [7]. This diagram used the data calculated from the data we collected. As shown in Fig 1, Turn Based games seem to have the highest average players but a really low average rating. There are some well-known turn based games like Civilization VI. Two possible reasons for this type of game having a low average rating could be: first, players might compare a turn based game with a well known and well designed one, like Civilization VI, and give the game a score lower than it actually deserved. Secondly, the success of some turn based games might born out many crude semi-finished products and botched imitators which only deserve a low score. Compared to the situation of turn based games, Pinball games went to the other extreme, high average score but low average players [8]. This is probably because there aren't many people are the fans of pinball games, and this little group of players are more accommodating to the quality of the games. On this diagram, we noticed that the performance of Shooter games is extremely terrible. This seemed surprising as there are so many successful shooter games like PUBG, Call of Duty series and Battlefield series. But it can still be explained, the reason is simple and had been mentioned before, there are too many botched imitators [9]. Thanks to powerful game engines, it became very easy to make a shooter game, but it is still hard to make it well. As a result, many badly designed shooter games are released, no one will give them good reviews and almost no one will play them. However, the existence of these terrible games has made shooter games performed badly on this graph. The coding for plotting this graph in Table 3

Table 3. The coding for plotting

```
df_graph1 = pd.DataFrame(np.zeros((53747,3)),columns = ['Genres','rating','total player count'])
count = 0
for i in range(0,len(Genres)):
   for j in range(0,60000):
      if Genres[i] in df_game.loc[j,'Genres']:
       if (df_game.loc[j,'Rating'] >= 0) :
            df_graph1.loc[count,'Genres'] = Genres[i]
            df_graph1.loc[count,'rating'] = df_game.loc[j,'Rating']
            df_graph1.loc[count,'total player count'] = df_game.loc[j,'Plays']
            count = count + 1
for i in range(0,53747):
   if str(df_graph1.loc[i,'total player count'])[len(str(df_graph1.loc[i,'total player count'])) - 1] == 'K':
      df_graph1.loc[i,'total player count'] = str(df_graph1.loc[i,'total player count'])[0:len(str(df_graph1.loc[i,'total player
count'])) - 1]
      df_graph1.loc[i,'total player count'] = float(df_graph1.loc[i,'total player count'])*1000
   df_graph1.loc[i,'total player count'] = float(df_graph1.loc[i,'total player count'])
df_graph1_rate = df_graph1.iloc[:,[0,1]]
df_graph1_number = df_graph1.iloc[:,[0,2]]
df_graph1_rate = df_graph1_rate.groupby('Genres').mean()
df_graph1_number = df_graph1_number.groupby('Genres').mean()
df_graph1 = pd.concat([df_graph1_rate,df_graph1_number],axis = 1)

graph1 = (ggplot(df_graph1,aes(x = 'total player count',
                    y = 'rating',
                    fill = 'Genres',
                    size = 10))
          +geom_point(
            shape = 'o',
            color = 'black',
            stroke = 0.2,
            alpha = 1
            )
          +theme(dpi = 300)
          +xlab ('average player')
          +ylab ('average rating')
          +geom_text(aes(label = 'Genres'),
                size = 10)
          )
print(graph1)
```

2.2 Number of games on each platforms

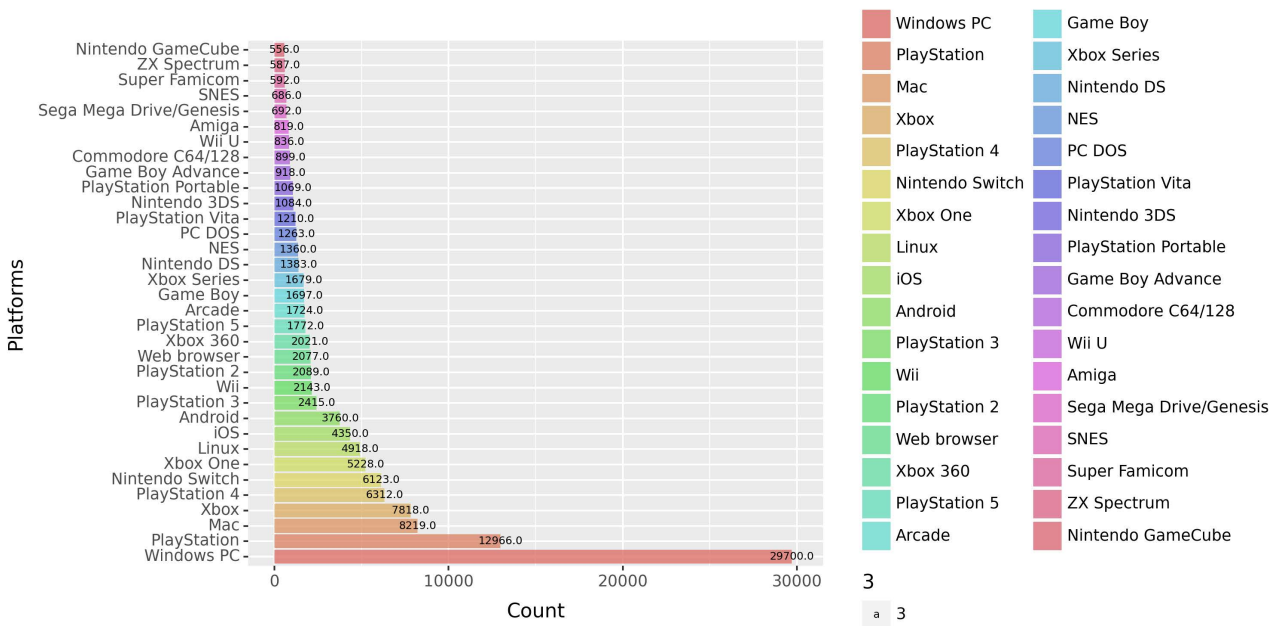

Fig. 2 Number of games on each platforms (Photo/Picture credit: Original).

As shown in Fig 2. There are over a hundred different platforms show up in the data we collected. But most of them are not popular at all [10]. So while plotting the graph showing number of games available on each platforms, we decided only to keep platforms with more than 500 games.

As shown in the bar chart above, those are all the platforms with over 500 games. In order to get the counting data, we wrote a program containing a loop which can look through all 60,000 game samples' platforms and count how many times each platforms have existed.

It is obviously that Windows PC is the platforms with the most games, it is then followed by PlayStation, Mac, Xbox, Nintendo Switch, iOS and Android. The reason for why there are so many games on PC is probably because most games are made in programming languages that support Windows PC. If a PC game want to make itself playable on another platform, it will have to fit itself in to the system of that platform. Another reason for the popularity of Windows PC is simple, because it got a very large number of users. However, it is actually not entirely a bad thing for other platforms. The difference in the system had also prevented games of low quality from spreading on multiple platforms. That is way games on Switch and PlayStation are usually fun to play. In the other words, even though Windows PC have many games, some of them are not good at all.

The coding for plotting in Table 4:

Table 4. The coding for plotting

```
platforms_count.info()
platforms_count['Platforms'] = platforms_count['Platforms'].astype(pd.CategoricalDtype(categories = platforms_count['Platforms'],ordered = False))
graph2 = (ggplot(platforms_count,aes(x = 'Platforms',
                                     y = 'Count',
                                     fill = 'Platforms'))
        +geom_bar(stat = 'identity',
                  alpha = 0.7)
        +geom_text(aes(x = 'Platforms',
                       y = 'Count',
                       label = 'Count',
                       size = 3),
                   nudge_y = 3)
        +theme(dpi = 300,
               figure_size = (10,5))
        +scale_y_continuous(limits=(0,30000))
        +coord_flip()
        )
graph2
```

2.3 Game Industry on the time scale

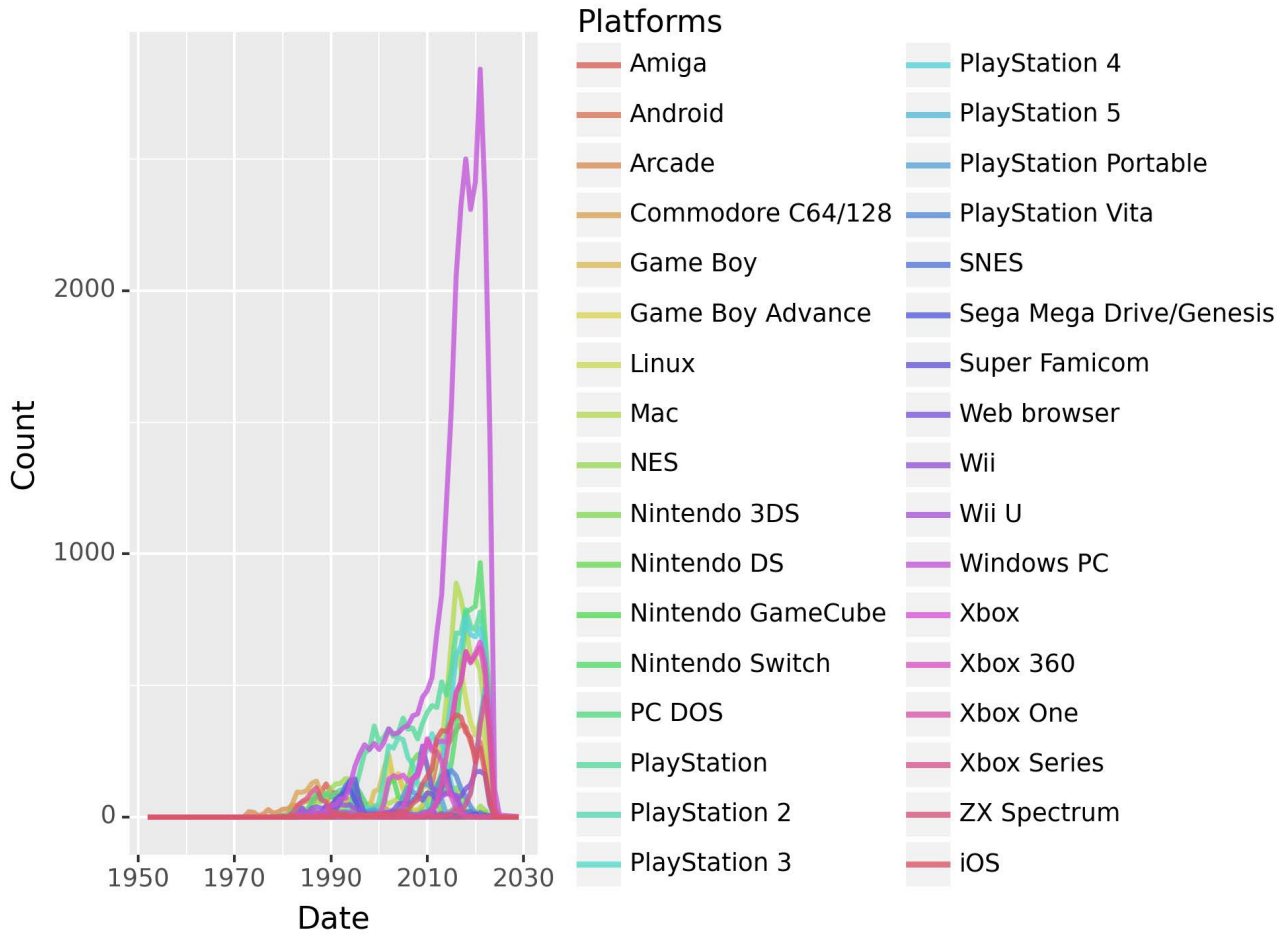

Fig. 3 Platforms on the time scale (Photo/Picture credit: Original).

As shown in fig 3, the line chart above showed information similar to the bar chart, the difference is that it has one more dimension, time. This line chart shows the number of games released on each platform in each year. We found that around the year 2000, PlayStation had similar number of games produced, compared to Windows PC. And after 2010, many games are released on Mac and Xbox had rose too. The number of games released all dropped back to zero after 2023 because there's no data in the future. Ignoring the drop after 2023, we saw a upward trend in lines of most of the platforms, Which means the game industry is still expending.

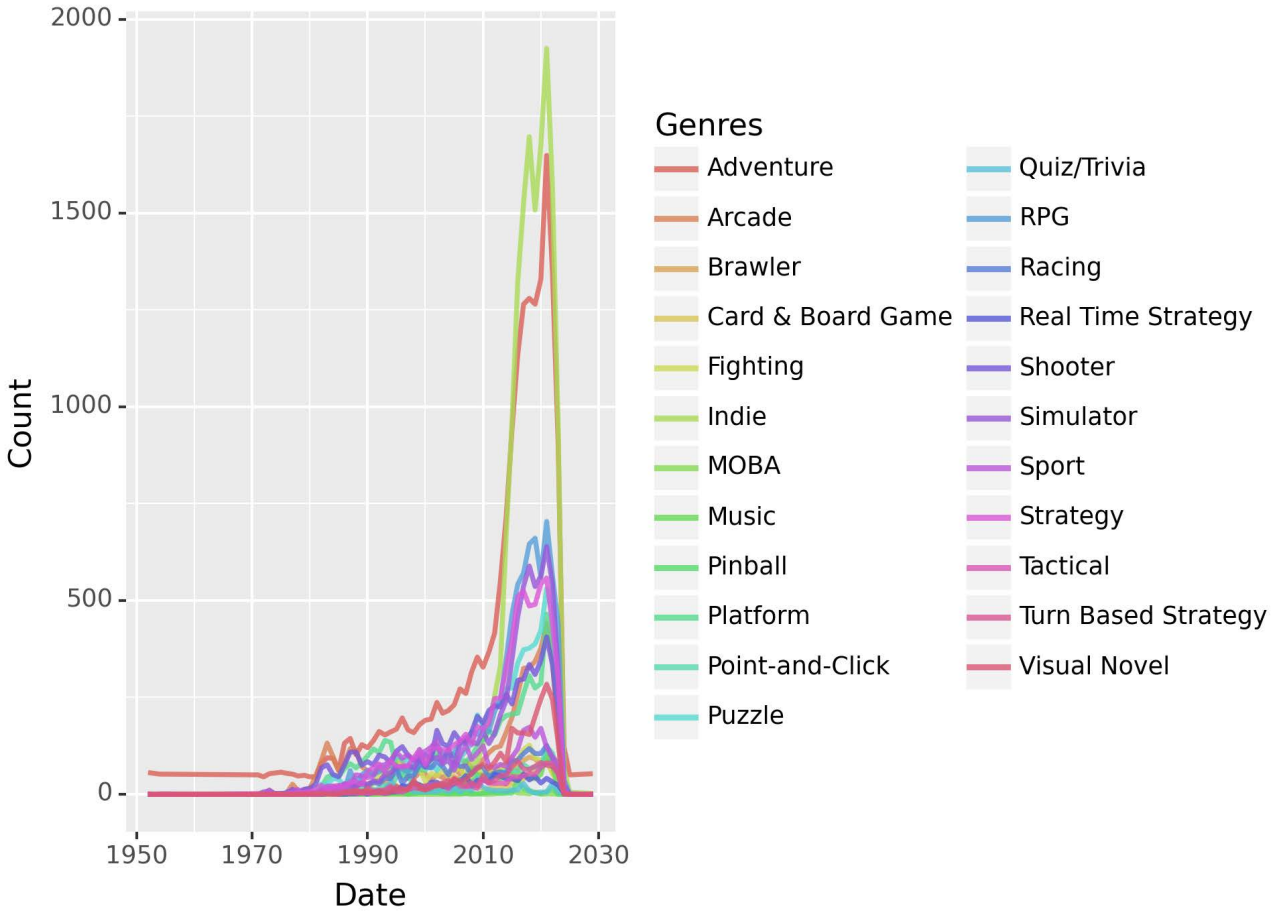

Fig. 4 Game genres on the time scale (Photo/Picture credit: Original).

As shown in Fig 4, this line chart showed number of games released in each year, but divided by different genres. As shown in the diagram, after 2010, Games with the labels 'Adventure' and 'Indie' have proliferated. And number of Indie games released even overcame that of Adventure games. An indie game means that the game is made by one or only a few people instead of companies. All these indie games, are probably made by those who grew up playing games. These game developers are possibly one of the players twenty years ago. They were impressed by video games and wanted to share their unique ideas to the other players. Apart from indie games, a lot of adventure games are released too. This could be explained by the taste of players. In adventure games, stories usually happen in a world completely different from the world we know and players could experience something that they could never get a chance to experience in real life. The coding for plotting in Table 5:

Table 5. The coding for plotting

```
df_graph3 = pd.DataFrame(np.zeros((100,23)),columns = Genres)
df_date = pd.DataFrame(np.zeros((100,1)),columns = ['Date'])
df_graph3 = pd.concat([df_graph3,df_date],axis = 1)
date = 0.0
count = 0
for i in range(0,60000):
  #for j in range(0,12):
      if ('2' in df_game.iloc[i,2]) or ('1' in df_game.iloc[i,2]):
         date = int(df_game.iloc[i,2][8:12])
         repetition = False
        for k in range(0,count):
           if date == df_graph3.iloc[k,23]:
```

```
                repetition = True
            if repetition == False:
                df_graph3.iloc[count,23] = date
                count = count + 1
df_graph3 = df_graph3.iloc[0:58,:]
df_graph3.to_csv('df_graph3.csv')
df_graph3 = pd.read_csv('df_graph3.csv')
df_graph3a = copy.deepcopy(df_graph3)
for i in range(0,60000):
    for j in range(0,58):
        if str(df_graph3a.loc[j,'Date'])[0:4] == df_game.iloc[i,2][8:12]:
            for k in range(0,len(Genres)):
                if (Genres[k] in df_game.iloc[i,6]) :
                    df_graph3a.iloc[j,k] = df_graph3a.iloc[j,k] + 1
''"
df_graph3b = df_graph3a.drop(columns = 'Pinball', inplace = True)
df_graph3a.columns = ['Adventure',
 'RPG',
 'Puzzle',
 'Brawler',
 'Indie',
 'Platform',
 'Turn Based Strategy',
 'Simulator',
 'Shooter',
 'Strategy',
 'Music',
 'Arcade',
 'Fighting',
 'Visual Novel',
 'Tactical',
 'Card & Board Game',
 'Sport',
 'Racing',
 'MOBA',
 'Point-and-Click',
 'Real Time Strategy',
 'Quiz/Trivia',
 'Pinball',
 'Date']
Genres
''"
df_graph3 = pd.read_csv('df_graph3.csv')
df_graph3a = pd.read_csv('df_graph3.csv')
df_graph3b = pd.melt(df_graph3a,id_vars = 'Date',var_name = 'Genres',value_name = 'Count')
graph3 = (ggplot(df_graph3b,aes(x = 'Date',
                        y = 'Count',
                        color = 'Genres'))
        +geom_line(size = 1,
                alpha = 0.8,
                )
        +theme(dpi = 300)
        )
graph3
graph3.save('Genres_date_count')
```

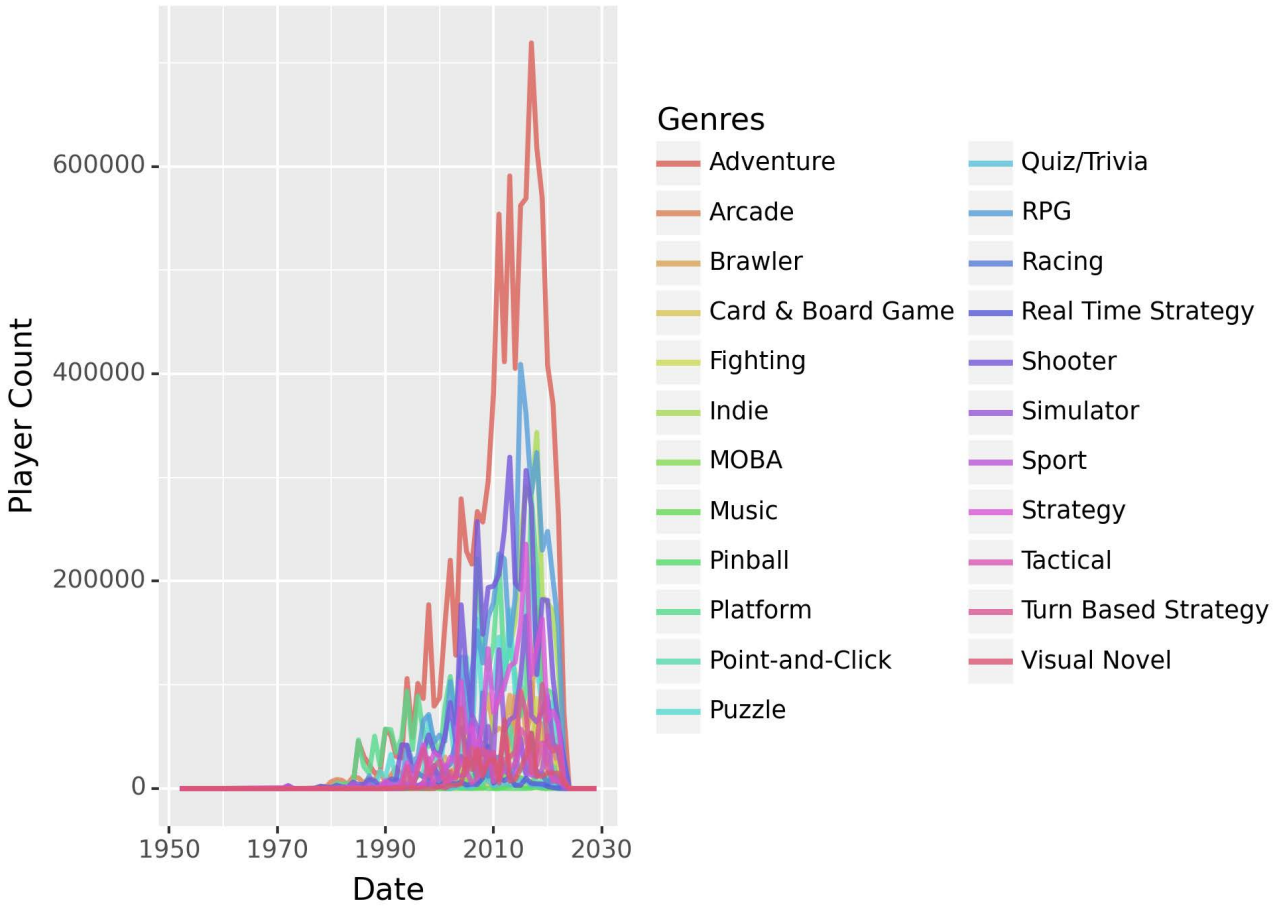

Fig. 5 Players count on different game genres on the time scale (Photo/Picture credit: Original).

As shown in Fig 5, the last line chart is a count on total number of players of different genres of games in each year. We can see that in the past twenty years, Adventure games are the most popular among all players, followed by RPG games, Indie games and shooter games.

3. Predictions

Now, here is some predictions about games.

If you want to make a game, the first thing to consider when making games is 'players'. As the direct service target of the game, players have a great influence on the game. The studio requires players to participate in the game and derive pleasure from any factors that would make them willing to persevere, so that the studio can gain profit from it, such as money, reputation, and even better technology. Therefore, the next question arises: How can we attract players?

One thing must be considered is the quality of the game. It determines whether the player will play this game. If the quality of the game is poor, players will not be willing to play a scrap. On the contrary, if the quality is good, players may play because they are interested or because the recommendation of others. Of course, there are exceptions. For example, although the quality of this game is not high, there are other very attractive aspects, such as the plot, fun, and playability. But at the same time, in order to have good game quality, employees must have strong skills. This also creates difficulties in finding employees. Also, Popularity can make big influence. The popularity of the game can vary in various aspects, such as the level of fun, the excitement of the plot, or the intense fighting. Any prominent area may attract more players. This tests the skills of game producers, who need to come up with some unique ideas to enable players to choose their own reasons from numerous games. But this is also the most difficult part of game design, because unlike game production, which requires time and effort, it requires more demanding imagination conditions. Next part is about players. It is hard to predict a player's interest. Because a person's interests can change at any time,

for various and even outrageous reasons. For example, 'gender'. A study did research on Singaporean, German, and American players to find game motives and genre due to gender. It shows that regarding gender differences in genre choices (RQ2), nearly all genres differed significantly by gender except strategy (marginally significant, $p = .06$) and platform games. Previous research suggests that game genres preference differs by gender, with some variation by country or region. In this study, men reported higher use of most genres, while women reported higher use than men for casual and music genres, which aligns with stereotypes of women gamers as focusing on less aggressive or competitive gaming experiences. At the same time, players have different points of interest. So if you want to create a game with more players, it is important to understand the overall interests of the players. From the charts above, we can see that most players enjoy playing large-scale games, with independent and adventure games being particularly prominent. As a result, game producers can lean towards these aspects to create games.

The viscosity of players is very important. It can keep players active in this game and continuously bring the game producers the results they want. What would happen if there was a little even no viscosity? Imagine that in this situation, fewer and fewer players are active in the game, and other players will also leave one after another, ultimately no one is playing. And the game requires funds and technology to support its operation. Without the benefits brought by players, the game cannot continue to run, and finally can only face the outcome of server suspension and removal or even no one's attention. This is a result that no game producer wants.

The 'influence' refers to both the player's impact on the game and the game's impact on the player. From the first perspective, a series of activities made by players both inside and outside the game will affect the development of the game, whether it is small or large. For example, players discovering bugs and providing feedback can help game producers fix them in a timely manner to improve game quality. For example, influential people like anchors who vigorously promote games can bring unimaginable popularity to this game. Of course, the impact may also be negative, but we will not expand on it.

Desire to consume may be something that most producers consider, not only because they need players to make money to make a profit, but also to allow their studios to establish themselves in this fiercely competitive market. Therefore, this is something that must be considered. In order to achieve this goal, game producers must set certain items or permissions in certain aspects of the game that only those who charge can use. For example, the buyout price in buyout games, the limited cards in turn based games, and the exquisite clothing in online games. Of course, many businesses in the market are using various methods to achieve this goal. But if there are too many activities that require recharging, it can have unimaginable consequences. It may make it difficult for an otherwise happy family to survive, and may even harm a country's economy.

4. Introspection

There are numerous variables not accounted for in our dataset, notably demographic factors like age, which could significantly influence the results. Additionally, the scarcity of studies with similar research objectives means there is limited empirical evidence to bolster our viewpoints. This lack of prior research is particularly evident in the predictive aspects of our analysis, where substantiating proofs are markedly sparse. These gaps highlight critical areas for reflection and underscore the need for a more comprehensive approach in future research. It is imperative to integrate a broader array of data points, including age and other relevant demographic information, to enhance the robustness of our predictions. Moreover, developing methodologies that can incorporate and analyze these variables will be crucial in advancing our understanding of the dynamics at play. The endeavor to expand the empirical foundation of our research will not only validate our findings but also pave the way for more informed and reliable predictive modeling in this field.

5. Conclusion

Creating a video game involves a myriad of considerations, making it a complex task for developers. Thus, meticulous planning and mental preparation are essential before initiating production. Beyond the game itself, it is crucial to consider external factors that might influence the project, such as advertising and marketing strategies. Moreover, while confidence in one's work is beneficial, game developers must temper this with a recognition of the inherent uncertainties of the future. The industry's dynamic nature means outcomes are unpredictable, and external trends or shifts in consumer preferences can dramatically impact a game's success. Consequently, game producers must focus on excelling in all aspects of their current responsibilities—ensuring that each element of development, from design to user testing, is executed to the highest standard. This approach not only prepares the game for potential success but also equips developers to adapt to challenges that may arise post-launch.

References

[1]Herrmann, M. R., Brumby, D. P., Oreszczyn, T., & Gilbert, X.

M. (2018). Does data visualization affect users' understanding of electricity consumption? Building Research & Information, 46(3), 238-250.

[2]Chen, C. W., & Hsu, T. (2018, April). Game development data analysis visualized with virtual reality. In 2018 IEEE International Conference on Applied System Invention (ICASI) (pp. 682-685). IEEE.

[3]YILDIRIM, İ. E., & ARIKAN KOKKAYA, G. Ü. L. E. N. (2022). Dimension reduction and visualization in big data: an analysis on the game industry.

[4]Zhu, X., Zhao, Z., Wei, X., & others. (2021). Action recognition method based on wavelet transform and neural network in wireless network. In 2021 5th International Conference on Digital Signal Processing (pp. 60-65).

[5]Copeland, B., Griffin, C., Bianco, C. E., Kononenko, N., & Craft, W. J. (2018). Data Visualization in Games. Major Qualifying Projects (All Years) (cit. on pp. 7, 14).

[6]Yampray, K., & Inchamnan, W. (2019, November). A method to visualization data collection by using gamification. In 2019 17th International Conference on ICT and Knowledge Engineering (ICT&KE) (pp. 1-4). IEEE.

[7]Bowman, B., Elmqvist, N., & Jankun-Kelly, T. J. (2012). Toward visualization for games: Theory, design space, and patterns. IEEE transactions on visualization and computer graphics, 18(11), 1956-1968.

[8]Drescher, C., Wallner, G., Kriglstein, S., Sifa, R., Drachen, A., & Pohl, M. (2018, April). What moves players? Visual data exploration of Twitter and gameplay data. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-13).

[9]Games, P. S., & Joshi, A. (2015, February). An evaluation-guided approach for effective data visualization on tablets. In Visualization and Data Analysis 2015 (Vol. 9397, pp. 8-20). SPIE.

[10]Pandey, A. V., Manivannan, A., Nov, O., Satterthwaite, M., & Bertini, E. (2014). The persuasive power of data visualization. IEEE transactions on visualization and computer graphics, 20(12), 2211-2220.