IMDB Video Games Data Analysis

Qingyue Tian

2025-06-28

Introduction and Background Information

This study aims to investigate the relationship between game ratings and sales, as well as the types, publishers, and platforms of games, between 2000 and 2020. For the ratings, I chose to use data from the large rating website IMDb because its ratings have a relatively large user base, the ratings are relatively reliable, and the dataset from its website can be found on Kaggle. I will present my research using various visual effects so that the audience can understand the content of my research most intuitively. Specifically, I will divide my research content into three sections. The first section examines the relationship between game ratings and sales. Is it true that the more people play the game, the higher the rating? This is what I care about. Can a good game in the traditional sense get the return it deserves in business? In the second section, we will delve deeper and add data such as the game's age, publisher, and platform to explore the characteristics of highly rated games in this era. In the third section, we will focus on one platform, the Wii platform, and examine the sales and ratings differences between first-party games and other third-party games on the platform. We will also investigate the impact of video game compatibility with the platform on their ratings.

About Our Data

My primary dataset is a dataset from Kaggle, which captures data from the IMDB rating website. Its credibility can be guaranteed by the credibility of the IMDB website. It is a dataset from three years ago, which can fully cover the data up to 2020. The extra two years will also make the data in 2020 more referenceable after time has passed. It includes the following attributes: name, url, year, certificate, rating, votes, plot, Action, Adventure, Comedy, Crime, Family, Fantasy, Mystery, Sci-Fi and Thriller. Among them, we will choose the name, year, rating, and votes columns to use. We will not use the genre column, which ranges from Action to Thriller, as the genre criterion. This is because, in the True and False criteria of these genre columns, as long as the game touches the edge of this category, it will be classified as True, which causes almost all games in the Action category to be classified as True, making the data

inconvenient to analyze. We only focus on works with more than 500 ratings.

My secondary datasets also come from Kaggle and Wikipedia. My first secondary dataset comes from Kaggle, which was uploaded two years ago by ULRIK THYGE PEDERSEN, a Kaggle-certified database expert. The upload time of this data is not significantly different from that of the previous data, so the two datasets can be compared and supplemented. The primary purpose of this dataset is to augment the first dataset with additional columns, including a Platform column, a Genre column, a Publisher column, and four sales columns. For the merger of dataset two and dataset one, since there are sufficient datasets and it is cumbersome to check and standardize the different names of each game, we will only select games with the same name in both datasets for analysis.

My second secondary data set comes from Wikipedia, which lists Wii games released. We primarily focus on its developers' column, from which we select Nintendo's first-party developed games and non-Nintendo developed games to compare and determine how the compatibility of games and platforms affects their ratings and sales.

Exploratory Data Analysis

First, we import the three data sets into the file and organize them, including selecting the required columns, filtering out rows whose data do not meet the requirements, and merging the data into a new dataset to facilitate our next step of visualizing the data.

Data Visualizations

Our first research topic is to consider the relationship between game sales and game ratings. Under this topic, we not only explore the simple relationship between total game sales and game ratings but also consider that IMDB is an American rating website, so sales in North America may be more critical to the rating than sales in other regions. We will use visual data to answer whether this hypothesis is true.

Rating vs Sales

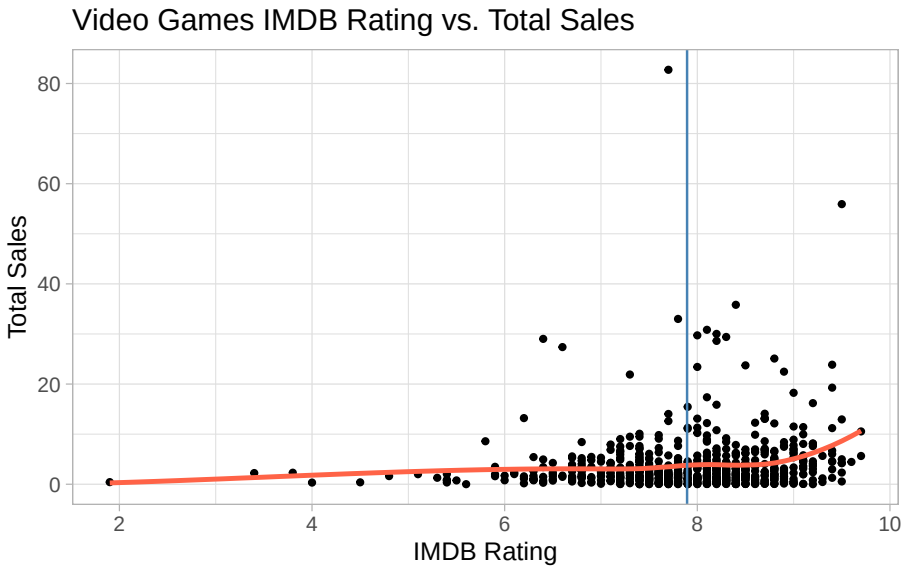

Figure 1

From the plot in Figure 1 , we can see that in the score range of 6 to 10, games with significantly high sales (sales exceeding 10 million) appear to be evenly distributed, with a slight concentration at a score of 8. The blue vertical line represents the average of all scores, as shown in Table 1, with a value of 7.9 points. The red line in the figure represents the trend line of this scatter plot. It can be seen from the trend chart that when the score exceeds 8.7 points, sales exhibit a significant upward trend compared to before. This shows that the score is indeed proportional to sales to a certain extent, and the sales of games with a score higher than a specific value (8.7 points) will also increase significantly. The good games considered by the public are indeed reflected in their commercial value.

Let's take a look at the visualization with North American data added.

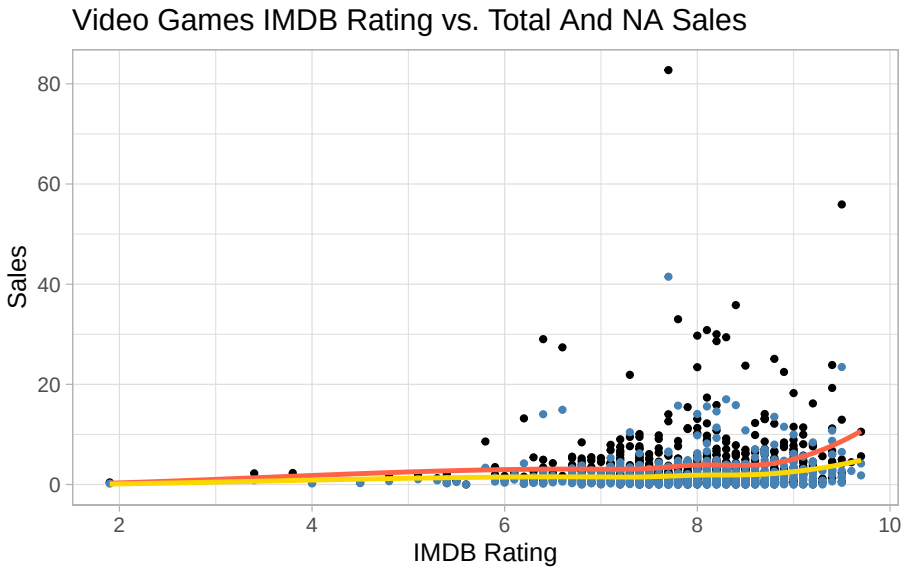

Figure 2

From the plot Figure 2, we can see that, compared to the total sales data, the North American sales data appears to be perfectly cut in half. This number seems to be inconsistent with the demographics. The total population of the United States and Canada is about 335 million, while the total population of Japan and Europe is about 865 million. This indicates that video game culture is more prevalent in North America than in other regions. In terms of the trend line, the general trend of the North American sales trend line is not significantly different from the total sales; however, the trend after the games with a score higher than 8.7 is relatively flat compared to the total sales. This indicates that games with higher scores will stimulate the purchasing desire of game enthusiasts worldwide.

It seems like North American sales are not more important than other countries in terms of ratings.

Our second research topic is whether the era, publisher, platform, and genre affect a game's rating data. In this section, we will explore whether a specific company is more successful with games of a particular genre.

Rating vs Published Year

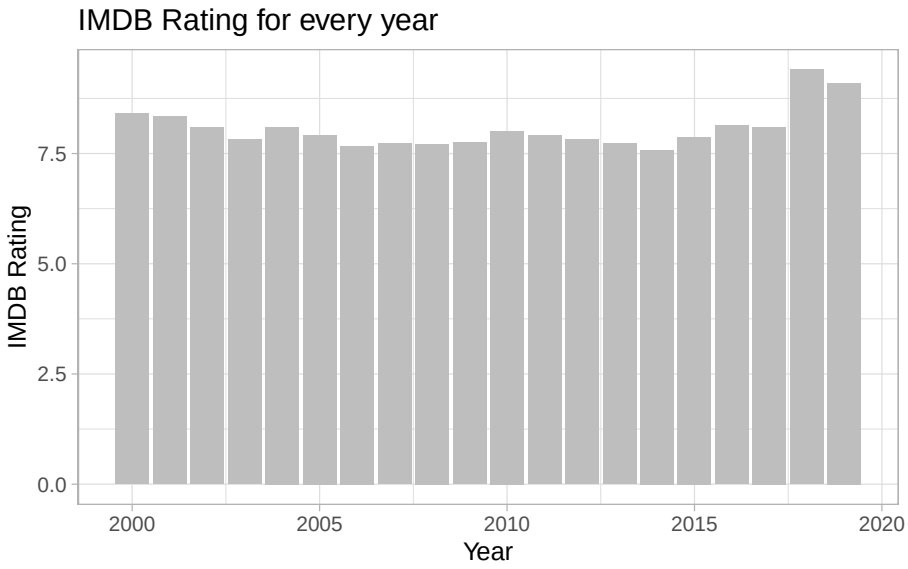

Figure 3

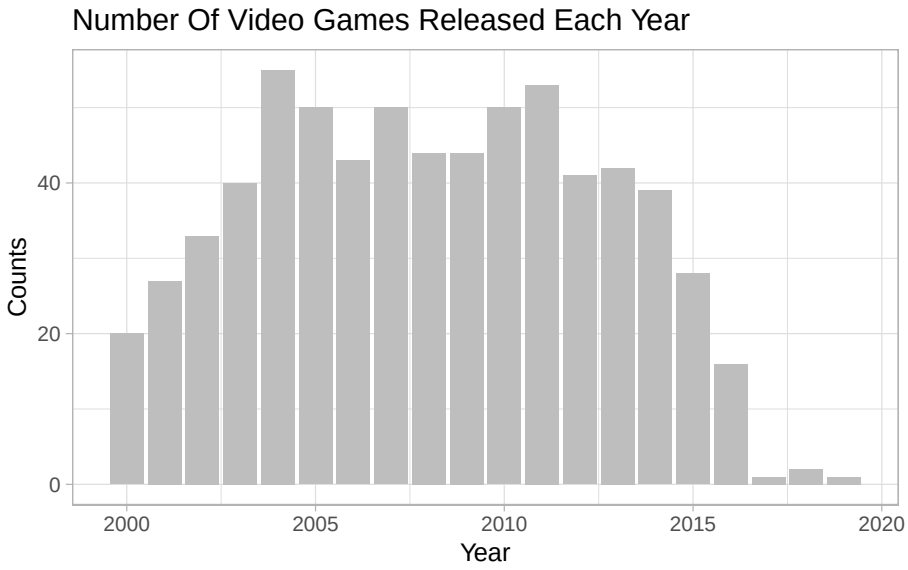

Figure 4

According to Figure 3, the ratings do not correlate with the year. In the second half of the figure, the difference between the data after 2006 and before is slightly larger because the amount of data has decreased significantly since 2017, making the average value look less average. I initially expected that in the release year of new game consoles, such as 2006 when the Wii was released, 2001 when the Xbox was released, and 2016 when the PlayStation 4 was released, the overall game ratings would have an upward trend because manufacturers would launch some big-budget games in conjunction with the consoles. However, the results in Figure 4 show that although there are specific significant differences in the number of game releases depending on the year, the release of a game console has no significant impact on the game's annual ratings and sales.

Since there are no significant fluctuations in game ratings in the years when game consoles are released, can we infer that game consoles (platforms) will not have a substantial impact on game ratings, and further, that the game manufacturers that manufacture game consoles (platforms) will not have a significant effect on game ratings?

Rating vs Platform

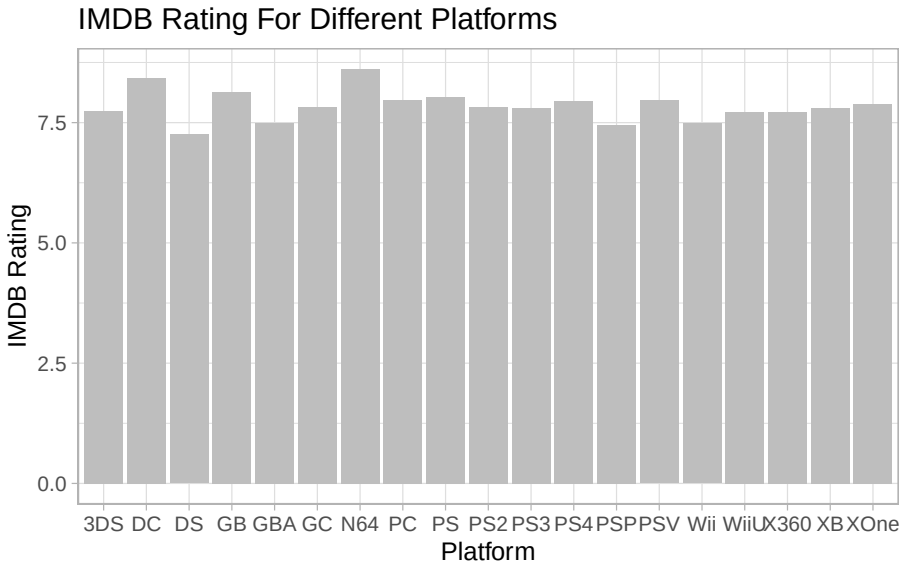

Figure 5

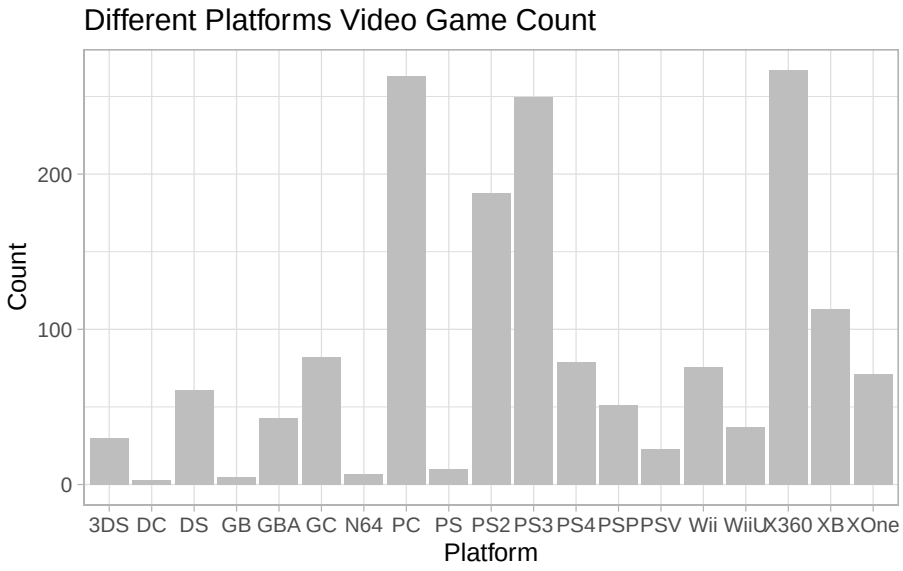

Figure 6

According to Figure 6 and Figure 5, although the number of games released on different platforms varies significantly, the difference in their average ratings is not substantial. The ratings of relatively popular platforms, such as PC, PS3, and X360, have not increased significantly compared to other ratings, but are closer to the average. Although games on different platforms have different characteristics, the image quality of PS4 is much better than that of PS2; the Wii, which must be connected to a TV, is significantly heavier than the 3DS, which is lightweight and can be played in the hand; however, the characteristics of lightness and image quality do not appear to impact the ratings of the game by players significantly.

Rating vs Publisher

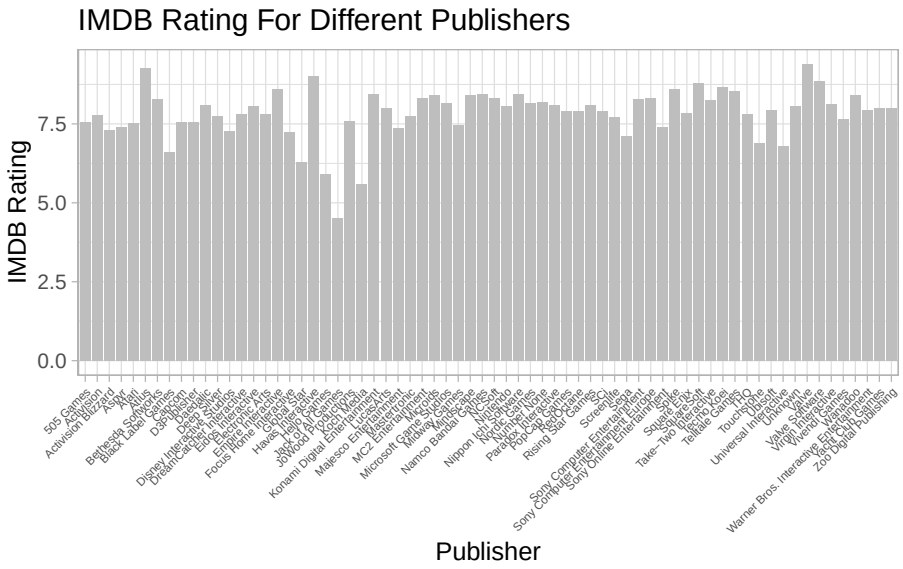

Figure 7

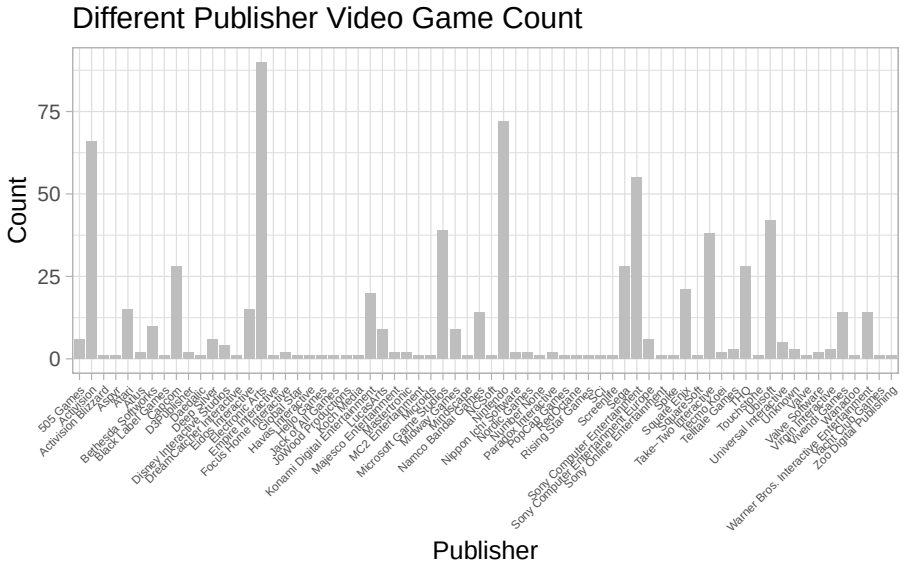

Figure 8

Table 1: The four publishers with the most games released

Publisher	count
Electronic Arts	90
Nintendo	72
Activision	66
Sony Computer Entertainment	55

From the above Figure 7 and Figure 8, we can see that the differences in scores between publishers are much greater than those between platforms. This is reasonable because it requires significantly more R&D investment to develop a platform than a game. From Figure 8, we can see that several large game manufacturers, such as Nintendo, Electronic Arts, Sony Computer Entertainment, and Ubisoft, all have scores of at least 7.5, indicating that the quality of games released by these manufacturers is generally high.

Now, let's turn our attention to the most basic attribute of the game: genre. Which games are sold the most in the market, and which games are likely to attract high ratings?s

Rating vs Genre

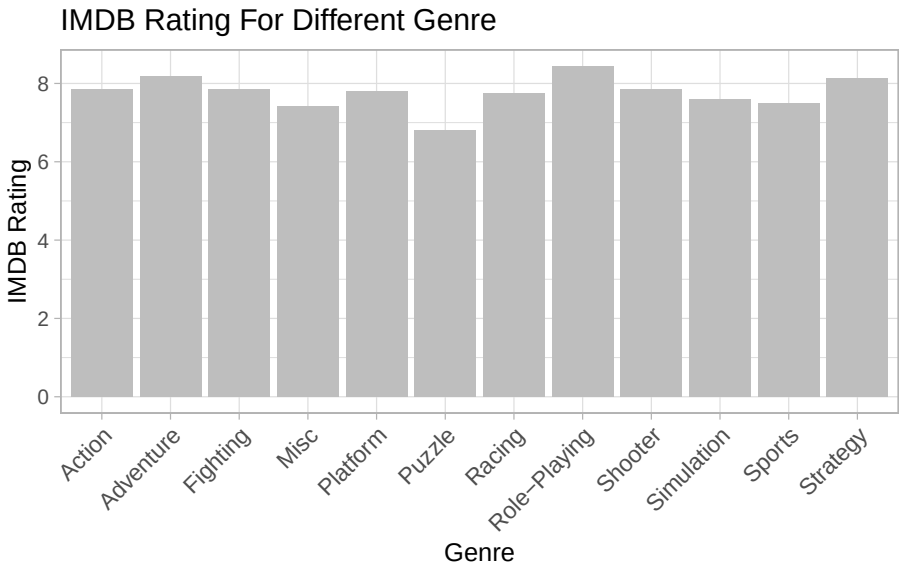

Figure 9

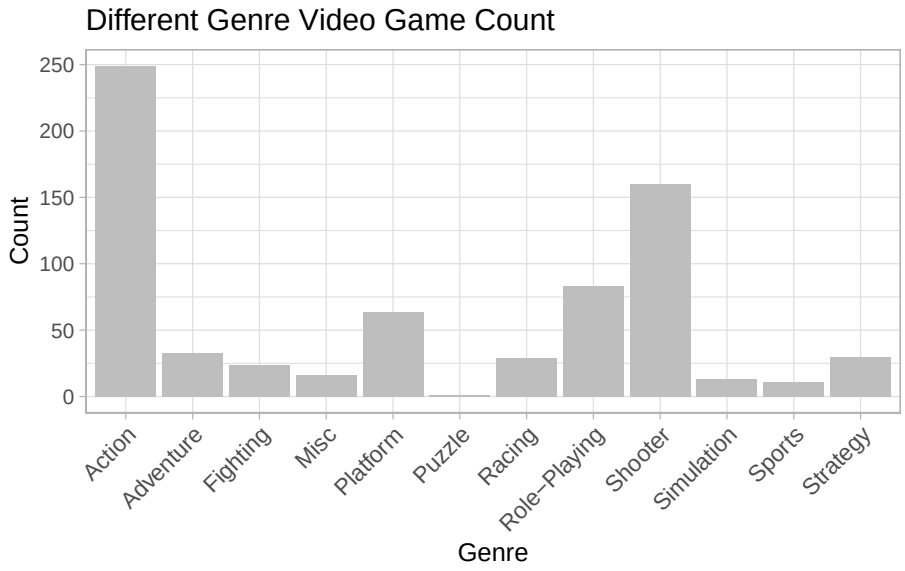

Figure 10

As can be seen from Figure 9 and Figure 10, the number of action games is significantly higher than that of other games. Design games are second, followed by role-playing games. In terms of ratings, role-playing games are the highest, followed by adventure games and strategy games. It is worth noting that puzzle games tend to have significantly lower sales and ratings compared to other types. As a game player, I can understand this phenomenon. Shooting games and action games have several "classic modes", such as horizontal action games and first-person shooter games, which are relatively easy for game manufacturers to reproduce. Role-playing games require a complete storyline, which makes their production somewhat challenging, but they are relatively easy to attract attention. After all, a successful character will make many groups that were not initially interested in it aware of its existence. For example, not many people in the world watch anime, but almost everyone is familiar with the existence of Doraemon.

As for adventure and strategy games, adventure games typically require more resources to create the stage of the adventure world during the production process, so there is a certain threshold for their release. This threshold also ensures their level to a certain extent. Strategy games focus on gameplay, and it is usually challenging to conceive new gameplay. However, good strategy games can be played almost infinitely, unlike adventure games or role-playing games with fixed clearance processes and time, so their scores will also be relatively high. Regarding puzzle games, their primary target audience is younger children, so they may not be as attractive to specific groups, especially those who

can consume and enter the website to score.

Nintendo

As shown in Figure 8, Nintendo is the second-largest game manufacturer in the data, releasing the second-most games between 2000 and 2020. At the same time, it also develops game consoles part-time. For example, Wii is a console designed by Nintendo. In Figure 6, you can see that our data includes more than 70 games. Next, we will narrow our horizons and focus on Nintendo's Wii consoles. By comparing Nintendo's first-party games with those of third-party developers, they will further explore the relationship between game consoles and ratings.

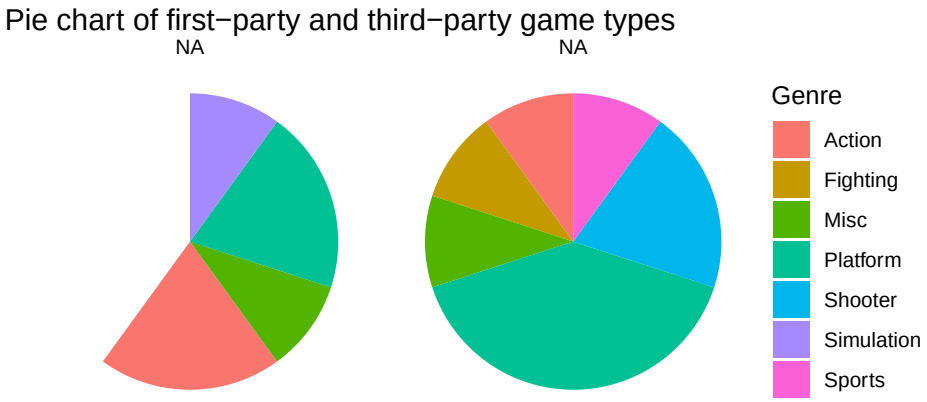

Figure 11

Table 2: A summary of the genres of first-party games and third-party games

Genre	Third-Party	First-Party
Action	1	2
Fighting	1	0
Misc	1	1
Platform	4	2
Shooter	2	0
Sports	1	0
Simulation	0	1

From Figure 11, we can see the types of first-party games and third-party games on the Wii. Third-party games tend to be more platform games and shooting games, while Nintendo's first-party games tend to be more platform and action games. From this, we can see that the Wii-type console is particularly well-suited for platform games, for example, the popular Super Monkey Ball series.

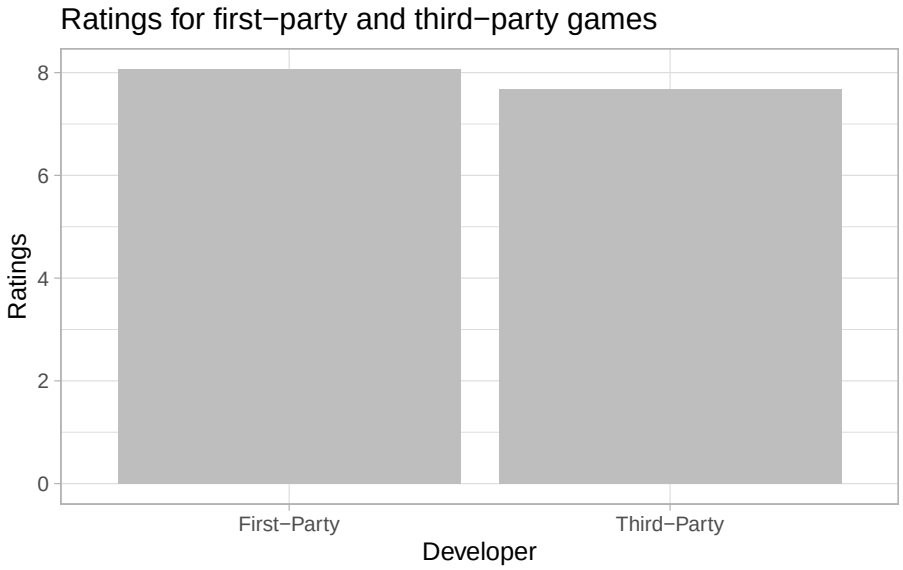

Figure 12

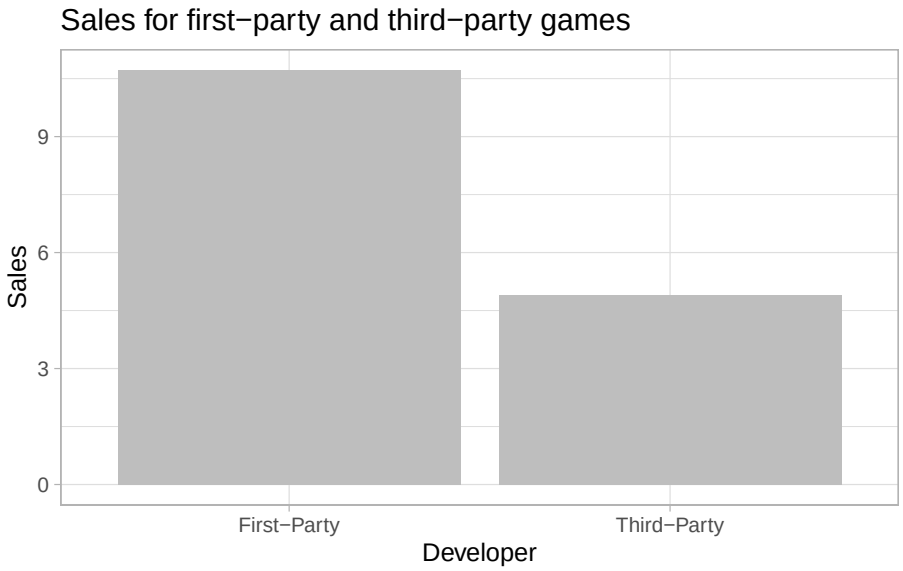

Figure 13

From Figure 12 and Figure 13, we can see that the average score of first-party games on the Nintendo Wii console is not significantly higher than that of third-party games. Still, the average sales of first-party games on the Nintendo console are considerably higher than those of third-party games. This shows that people are still more willing to buy first-party games that are compatible with the console, and their quality is guaranteed.

Conclusion

The analysis of video games from 2000 to 2020 helps us gain a better understanding of the game market during that period. From the game sales records and the ratings on the IMDb website, we can conclude that, although it is not immediately apparent, the game rating is proportional to sales, and the sales of games with a score of 9 or above exhibit a clear upward trend. The year of the game does not significantly impact the game rating, which suggests that whether it is an old, pixel-style game or a new 3A masterpiece, there will be a group of gamers who appreciate this aspect. The playing method, such as handhelds like the 3DS, consoles like the PS4, and PC, has little impact on the game rating. We found that game companies tend to develop shooting games and action games, while games with higher ratings are often role-playing games and adventure games. We found that games released by large game companies typically come with a high-quality guarantee. Finally, by comparing the data of first-party games and third-party games for Nintendo Wii, we found that first-party games are similar to third-party games in terms of ratings. Still, they

are far better than third-party games in terms of sales, which suggests that a certain segment of gamers is more inclined to buy first-party games.

Reference

Wikimedia Foundation. (2025, May 22). *List of best-selling Wii Video Games*. Wikipedia. https://en.wikipedia.org/wiki/List_of_best-selling_Wii_video_games

Feroze, Z. (2025, March 21). *Video games sale*. Kaggle.

https://www.kaggle.com/datasets/zahidmughal2343/video-games-sale

Talay, M. A. (2022, September 2). *IMDB video games*. Kaggle.

https://www.kaggle.com/datasets/muhammadadiltalay/imdb-video-games

Code Appendix

```r
# Load necessary packages
# The code in this pdf follows Google's R Style Guide
library(ggplot2)
library(dplyr)
library(tidyverse)
library(tidyr)
library(rvest)
library(readr)
library(knitr)

#Importing the Data
IMDBDataSetRaw <- read.csv(
  "/Users/maming/Downloads/imdb-videogames.csv"
  )
SalesDataSetRaw <- read.csv(
  "/Users/maming/Downloads/video games sales.csv"
  )
WebsitePage <-
  "https://en.wikipedia.org/wiki/List_of_best-selling_Wii_video_games"
TableList <- WebsitePage %>%
  read_html() %>%
  html_nodes(css = "table") %>%
  html_table(fill = TRUE)
WiiDataSetRaw <- TableList[[3]]

#cleaning the three raw data
IMDBCleaned <- IMDBDataSetRaw %>%
  filter(year >= 2000 & year <= 2020) %>%
  select(name, year, rating, votes) %>%
  mutate(votes = gsub(",", "", votes)) %>%
  mutate(votes = as.integer(votes)) %>%
  filter(votes >= 500) %>%
  filter(!is.na(rating) & rating != "")

SalesCleaned <- SalesDataSetRaw %>%
  filter(Year >= 2000 & Year <= 2020) %>%
  filter(!is.na(Platform) & Platform != "") %>%
  filter(!is.na(Genre) & Genre != "") %>%
  filter(!is.na(Publisher) & Publisher != "") %>%
  filter(!is.na(NA_Sales) & NA_Sales != "") %>%
  filter(!is.na(EU_Sales) & EU_Sales != "") %>%
  filter(!is.na(JP_Sales) & JP_Sales != "") %>%
  filter(!is.na(Other_Sales) & Other_Sales != "") %>%
```

```r
  filter(!is.na(Global_Sales) & Global_Sales != "") %>%
  select(-Rank)

WiiCleaned <- WiiDataSetRaw %>%
  mutate(Game = gsub("†", "", Game)) %>%
  mutate(Game = gsub("\\[e\\]", "", Game)) %>%
  rename(releaseDate = "Release date[a]",
         Developer = "Developer(s)", Name = Game) %>%
  mutate(releaseDate = format(as.Date(
    releaseDate, format = "%B %d, %Y"), "%Y")) %>%
  mutate(Nintendo = grepl("Nintendo EAD", Developer)) %>%
  select(Name, releaseDate, Nintendo)

#Merge data frames into new data frames for easy visualization
SalesMerge <- SalesCleaned %>%
  select(Name, Global_Sales, JP_Sales,
         NA_Sales, EU_Sales, Other_Sales) %>%
  group_by(Name) %>%
  summarise(
    TotalSales = sum(Global_Sales, na.rm = TRUE),
    JPSsale = sum(JP_Sales, na.rm = TRUE),
    NASales = sum(NA_Sales, na.rm = TRUE),
    EUSales = sum(EU_Sales, na.rm = TRUE),
    OtherSales = sum(Other_Sales, na.rm = TRUE),
    .groups = "drop"
  )
IMDBMerge <- IMDBCleaned %>%
  group_by(name) %>%
  slice_max(votes, with_ties = FALSE) %>%
  ungroup()

# Data frames for rating vs sales
RatingSales <- inner_join(
  IMDBMerge, SalesMerge, by = c("name" = "Name"))
AverageRating <- RatingSales %>%
  summarize(avgRating = mean(rating))

# Data frames for rating vs year
RatingYear <- RatingSales %>%
  select(year, rating) %>%
  group_by(year) %>%
  summarize(RatY = mean(rating))
AmontVG <- RatingSales %>%
  select(year, rating) %>%
  group_by(year) %>%
```

```
  summarise(count = n())

# Data frames for rating vs platform
SalesMergePlatform <- SalesCleaned %>%
  select(Name, Platform, Genre, Publisher, Global_Sales)
RatingPlatformMerge <- inner_join(
  IMDBMerge, SalesMergePlatform, by = c("name" = "Name"))
RatingPlatform <- RatingPlatformMerge %>%
  select(Platform, rating) %>%
  group_by(Platform) %>%
  summarize(RatP = mean(rating))
RatingPlatformCount <- RatingPlatformMerge %>%
  select(Platform, rating) %>%
  group_by(Platform) %>%
  summarise(count = n())

# Data frames for rating vs publishers
SalesMergePG <- SalesCleaned %>%
  select(Name, Genre, Publisher) %>%
  distinct()
RatingPGMerge <- inner_join(
  IMDBMerge, SalesMergePG, by = c("name" = "Name"))
RatingPublisher <- RatingPGMerge %>%
  select(Publisher, rating) %>%
  group_by(Publisher) %>%
  summarize(RatPub = mean(rating))
RatingPublisherCount <- RatingPGMerge %>%
  select(Publisher, rating) %>%
  group_by(Publisher) %>%
  summarise(count = n())

# Data frames for rating vs genre
RatingGenre <- RatingPGMerge %>%
  select(Genre, rating) %>%
  group_by(Genre) %>%
  summarize(RatG = mean(rating))
RatingGenreCount <- RatingPGMerge %>%
  select(Genre, rating) %>%
  group_by(Genre) %>%
  summarise(count = n())

#Nintendo Wii data frames
SalesWii <- SalesCleaned %>%
  filter(Platform == "Wii") %>%
  select(Name, Genre, Global_Sales)
```

```r
RatingWii <- inner_join(
  IMDBMerge, SalesWii, by = c("name" = "Name"))
RatingWiiN <- inner_join(
  RatingWii, WiiCleaned, by = c("name" = "Name"))
RatingWiiNG <- RatingWiiN %>%
  group_by(Nintendo, Genre) %>%
  summarise(count = n(), .groups = "drop") %>%
  mutate(Nintendo = factor(Nintendo,
                           levels = c("TRUE", "FALSE"),
                           labels = c("First-Party", "Third-Party")))
RatingWiiNG_wide <- RatingWiiNG %>%
  pivot_wider(
    names_from = Nintendo,
    values_from = count,
    values_fill = 0
  )
RatingWiiNR <- RatingWiiN %>%
  group_by(Nintendo) %>%
  summarize(RatN = mean(rating), .groups = "drop") %>%
  mutate(Nintendo = factor(Nintendo,
                           levels = c("TRUE", "FALSE"),
                           labels = c("First-Party", "Third-Party")))
RatingWiiNS <- RatingWiiN %>%
  group_by(Nintendo) %>%
  summarize(SalN = mean(Global_Sales), .groups = "drop") %>%
  mutate(Nintendo = factor(Nintendo,
                           levels = c("TRUE", "FALSE"),
                           labels = c("First-Party", "Third-Party")))

# Visualization of video game total sales and IMDB ratings
ggplot()+
  geom_point(data = RatingSales,
             mapping = aes(x = rating, y = TotalSales), size = 1)+
  geom_smooth(data = RatingSales,
              mapping = aes(x = rating, y = TotalSales),
              method = "loess", color = "tomato", se = FALSE)+
  geom_vline(data = AverageRating,
             aes(xintercept = avgRating), color = "steelblue")+
  labs(
    x = "IMDB Rating",
    y = "Total Sales",
    title = "Video Games IMDB Rating vs. Total Sales"
  )+
theme_light()

```

```r
# Visualization of video game north american sales and IMDB ratings
ggplot( data = RatingSales, )+
  geom_point(mapping = aes(x = rating, y = TotalSales),
             size = 1)+
  geom_point(mapping = aes(x = rating, y = NASales),
             size = 1, color = "steelblue")+
  geom_smooth(data = RatingSales,
              mapping = aes(x = rating, y = TotalSales),
              method = "loess", color = "tomato", se = FALSE)+
  geom_smooth(data = RatingSales,
              mapping = aes(x = rating, y = NASales),
              method = "loess", color = "gold", se = FALSE)+
  labs(
    x = "IMDB Rating",
    y = "Sales",
    title = "Video Games IMDB Rating vs. Total And NA Sales"
  )+
theme_light()

# Visualization of video games published year and IMDB ratings
ggplot(RatingYear, aes(x = year, y = RatY)) +
  geom_col(fill = "grey") +
  theme_minimal() +
  labs(title = "IMDB Rating for every year",
       x = "Year",
       y = "IMDB Rating") +
  theme_light()

# Visualization of number of video games released by year
ggplot(AmontVG, aes(x = year, y = count)) +
  geom_col(fill = "grey") +
  theme_minimal() +
  labs(title = "Number Of Video Games Released Each Year",
       x = "Year",
       y = "Counts") +
  theme_light()

# Visualization of video games on different plantform and their IMDB ratings
ggplot(RatingPlatform, aes(x = Platform, y = RatP)) +
  geom_col(fill = "grey") +
  theme_minimal() +
  labs(title = "IMDB Rating For Different Platforms",
       x = "Platform",
       y = "IMDB Rating") +
  theme_light()
```

```r

# Visualization of number of video games on different plantform
ggplot(RatingPlatformCount, aes(x = Platform, y = count)) +
  geom_col(fill = "grey") +
  theme_minimal() +
  labs(title = "Different Platforms Video Game Count",
       x = "Platform",
       y = "Count") +
  theme_light()

# Visualization of video games on different publishers and their IMDB ratings
ggplot(RatingPublisher, aes(x = Publisher, y = RatPub)) +
  geom_col(fill = "grey") +
  theme_minimal() +
  labs(title = "IMDB Rating For Different Publishers",
       x = "Publisher",
       y = "IMDB Rating") +
  theme_light() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1,
                                   size = 5))

# Visualization of number of video games on different publisher
ggplot(RatingPublisherCount, aes(x = Publisher, y = count)) +
  geom_col(fill = "grey") +
  theme_minimal() +
  labs(title = "Different Publisher Video Game Count",
       x = "Publisher",
       y = "Count") +
  theme_light() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1,
                                   size = 5))

# Table of number of video games on different publisher
RatingPublisherCount %>%
  arrange(desc(count)) %>%
  slice_head(n = 4) %>%
  kable(caption = "The four publishers with the most games released")

# Visualization of video games on different genre and their IMDB ratings
ggplot(RatingGenre, aes(x = Genre, y = RatG)) +
  geom_col(fill = "grey") +
  theme_minimal() +
  labs(title = "IMDB Rating For Different Genre",
       x = "Genre",
       y = "IMDB Rating") +
```

```
  theme_light() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1,
                                   size = 10))

# Visualization of number of video games on different genre
ggplot(RatingGenreCount, aes(x = Genre, y = count)) +
  geom_col(fill = "grey") +
  theme_minimal() +
  labs(title = "Different Genre Video Game Count",
       x = "Genre",
       y = "Count") +
  theme_light() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1,
                                   size = 10))

# Visualization of video games on Nintendo Wii Genre by developer
ggplot(RatingWiiNG, aes(x = "", y = count, fill = Genre)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  facet_wrap(~ Nintendo, labeller = as_labeller(c(
    "TRUE" = "First-Party",
    "FALSE" = "Third-Party"))) +
  theme_void() +
  labs(title = "Pie chart of first-party and third-party game types")

# A table of the genres of first-party games and third-party games
kable(RatingWiiNG_wide,
      caption = "A summary of the genres of first-party games and third-party games")

# Visualization of video games on Nintendo Wii Rating by developer
ggplot(RatingWiiNR, aes(x = Nintendo, y = RatN)) +
  geom_col(fill = "grey") +
  theme_minimal() +
  labs(title = "Ratings for first-party and third-party games",
       x = "Developer",
       y = "Ratings") +
  theme_light()

# Visualization of video games on Nintendo Wii sales by developer
ggplot(RatingWiiNS, aes(x = Nintendo, y = SalN)) +
  geom_col(fill = "grey") +
  theme_minimal() +
  labs(title = "Sales for first-party and third-party games",
       x = "Developer",
       y = "Sales") +
```

```
314   theme_light()
```