# Video game sales trend and data visualization

2023-05-15

## Contents



### Data Origin

The source of the original data is Kaggle https://www.kaggle.com/datasets/gregorut/videogamesales.The Video Game Sales dataset contains information on video games that have sold over 100,000 copies between 1980 and 2016. The data is generated by scraping https://www.vgchartz.comand includes the game's name, release date, platform, genre, publisher, sales figures in major regions, and total global sales figures.

### Research Question

The development of electronic games is always an interesting topic, especially between the 20th and 21st centuries. The upgrades in electronic hardware and software development have rapidly spread the video game entertainment category worldwide.

My first aim is to use data visualization to understand whether video game sales vary with the number of game releases, or in other words, whether the relationship between total annual sales and the number of game releases is stable. Or is it more likely that changes in sales are generally influenced by other larger factors (such as a few classic masterpiece games that greatly boosted revenue in a given year, the popularity and updates of gaming platforms, etc.)?

Second, I want to explore the top ten global best-selling game publishers and their market share in the entire gaming market. This will help us understand the composition of the video gaming market over several decades.

### Preparation for Data

Due to the importance of the release year and publisher data, all rows with N/A values in the year and publisher columns was removed. In addition, two years with significantly insufficient data (year=2017, 2020) were also excluded. To create the visualizations, make sure to load the following packages:

```
library(tidyverse)
library(here)
library(ggplot2)
library(plotly)
library(RColorBrewer)
```

```
#load the data set
vgsales<-read.csv(here("data","vgsales.csv"))
#show first few rows of the raw data set
head(vgsales)
```

```
##   Rank                     Name Platform Year        Genre Publisher NA_Sales
## 1    1               Wii Sports      Wii 2006       Sports  Nintendo    41.49
## 2    2        Super Mario Bros.      NES 1985     Platform  Nintendo    29.08
## 3    3           Mario Kart Wii      Wii 2008       Racing  Nintendo    15.85
## 4    4        Wii Sports Resort      Wii 2009       Sports  Nintendo    15.75
## 5    5 Pokemon Red/Pokemon Blue       GB 1996 Role-Playing  Nintendo    11.27
## 6    6                   Tetris       GB 1989       Puzzle  Nintendo    23.20
##   EU_Sales JP_Sales Other_Sales Global_Sales
## 1    29.02     3.77        8.46        82.74
## 2     3.58     6.81        0.77        40.24
## 3    12.88     3.79        3.31        35.82
## 4    11.01     3.28        2.96        33.00
## 5     8.89    10.22        1.00        31.37
## 6     2.26     4.22        0.58        30.26
```

```
#Clean up important missing rows in raw data
vgsales <- vgsales[vgsales$Year!='N/A',]
vgsales <- vgsales[vgsales$Publisher!='N/A',]
#Excluding the two most recent years with apparent data incompleteness
vgsales<-vgsales[!(vgsales$Year%in% c(2020, 2017)), ]
#check the data
head(vgsales)
```

```
##   Rank                     Name Platform Year        Genre Publisher NA_Sales
## 1    1               Wii Sports      Wii 2006       Sports  Nintendo    41.49
## 2    2        Super Mario Bros.      NES 1985     Platform  Nintendo    29.08
## 3    3           Mario Kart Wii      Wii 2008       Racing  Nintendo    15.85
## 4    4        Wii Sports Resort      Wii 2009       Sports  Nintendo    15.75
## 5    5 Pokemon Red/Pokemon Blue       GB 1996 Role-Playing  Nintendo    11.27
## 6    6                   Tetris       GB 1989       Puzzle  Nintendo    23.20
##   EU_Sales JP_Sales Other_Sales Global_Sales
## 1    29.02     3.77        8.46        82.74
## 2     3.58     6.81        0.77        40.24
## 3    12.88     3.79        3.31        35.82
## 4    11.01     3.28        2.96        33.00
## 5     8.89    10.22        1.00        31.37
## 6     2.26     4.22        0.58        30.26
```

### Preparing Data for Visualisation 1

```
#Group and sum data by year and sales
sales_by_year <- vgsales %>%
  group_by(Year) %>%
  summarize(total_sales = sum(Global_Sales))
#Count the number of games by year
games_by_year <- vgsales %>%
  group_by(Year) %>%
  summarize(num_games = n())
#Merge the data to get a new dataset with release year, release number and their sales by year
total_sales_by_year <- merge(sales_by_year, games_by_year, by = "Year")
#Check the data
head(total_sales_by_year)
```

```
##   Year total_sales num_games
## 1 1980       11.38         9
## 2 1981       35.77        46
## 3 1982       28.86        36
## 4 1983       16.79        17
## 5 1984       50.36        14
## 6 1985       53.94        14
```

### Preparing Data for Visualisation 2

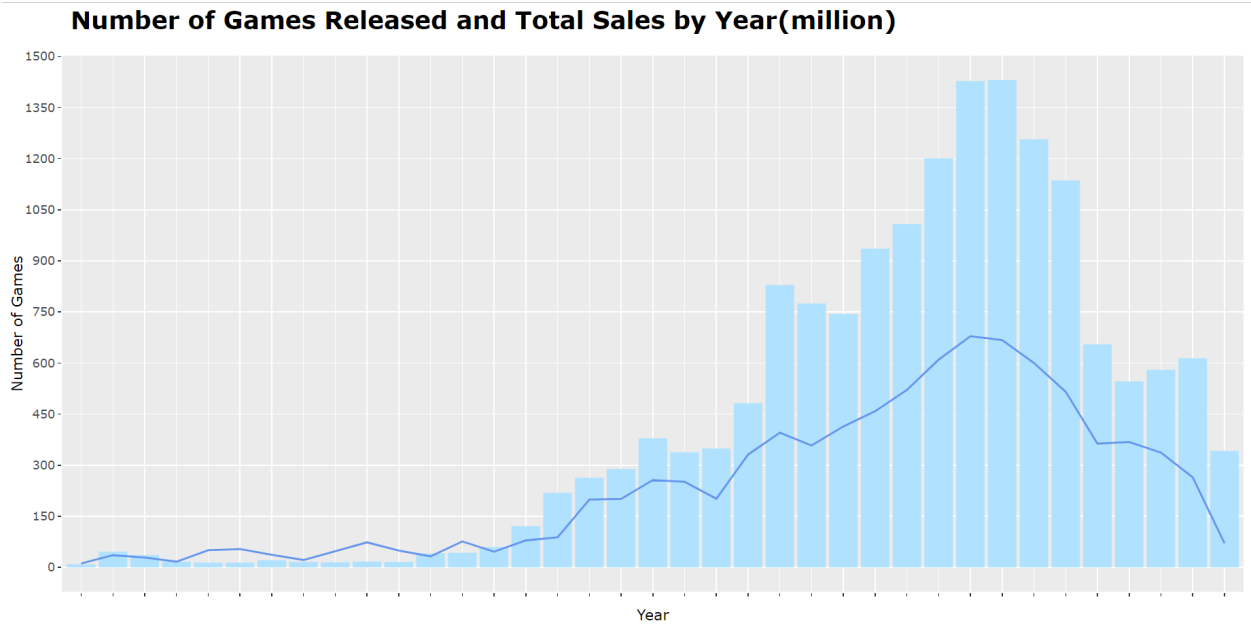
```
#Extract the top ten publishers by sales and their global sales
Top10_Pub <-vgsales %>%
  group_by(Publisher) %>%
  summarize(total_sales = sum(Global_Sales)) %>%
  top_n(10, total_sales) %>%
  arrange(desc(total_sales))
#Extract global sales of other publishers and combine into one
Other_Pub <- vgsales %>%
  group_by(Publisher) %>%
  summarize(total_sales = sum(Global_Sales)) %>%
  filter(!Publisher %in% Top10_Pub$Publisher) %>%
  summarize(Publisher = "other", total_sales = sum(total_sales))
#Merge the data of the top ten publishers and other publishers
Global_Pub <- bind_rows(Top10_Pub, Other_Pub)
#Check the data
head(Global_Pub)
```

```
## # A tibble: 6 x 2
##   Publisher                   total_sales
##   <chr>                             <dbl>
## 1 Nintendo                          1784.
## 2 Electronic Arts                   1093.
## 3 Activision                         721.
## 4 Sony Computer Entertainment        607.
## 5 Ubisoft                            473.
## 6 Take-Two Interactive               399.
```

## Visualisation 1

This data is visualized using a bar chart with a line, which compared to overlaid histograms, the trend of the data in the image clearer and also solves the problem of inconsistent intervals between the two y-axes.

```
#use ggplot to create a bar chart with line
Release_Sales<-ggplot(total_sales_by_year, aes(x = Year)) +
  geom_bar(aes(y = num_games), stat = "identity", fill = "lightskyblue1") +
  geom_line(aes(y = total_sales, group = 1), color = "cornflowerblue") +
  #Add a second y axis
  scale_y_continuous(breaks=seq(0,1500,by=150),sec.axis = sec_axis(~./10, name = "Total Sales (millions)
  labs(x = "Year", y = "Number of Games",
       title = "Number of Games Released and Total Sales by Year(million)")+
  theme(plot.title = element_text(size = 13, face = "bold"),
        #cancelling the display of the year for the sake of beauty
        axis.text.x=element_blank())
#Use plotly to make the year&sales display on mouse hover
ggplotly(Release_Sales, tooltip = c("x", "total_sales"))
```




### Summary of Visualisation 1

By visualizing the data, it is evident that in most of the time, the relationship between sales and release numbers is stable, meaning that the total sales of video games are usually positively correlated with the number of releases. However, we can also see that there are some years where the number of releases is negatively correlated with total sales. I speculate that this could be due to changes in game quality (as previously mentioned regarding the emergence of masterpiece games) or fluctuations in pricing (annual game prices being much higher or lower than the average).

In addition, I also noticed that there was a significant growth in data from 2002, it possibly due to the movement towards networked video game platforms and the enormous development that followed. However, from 2012, there was a noticeable decline in the activity of the video game market. Interestingly, the reason for this change may not be due to any problem or innovation in video games themselves, but rather because the entire gaming market began to gradually shift towards smartphones from this period.(source: Wikipediahttps://en.wikipedia.org/wiki/Video_game.)

### Visualisation 2

This data is visualized using a pie chart, which provides a more intuitive display of the percentage of each category compared to other charts.

```
ggplot(Global_Pub, aes(x = "", y = total_sales, fill = Publisher)) +
  geom_bar(stat = "identity",color="white") +
  coord_polar("y", start = 0) +
  #settin color of pie chart
  scale_fill_brewer(palette = "RdYlGn") +
  theme_classic() +
  #Hide unnecessary elements in the chart
  theme(axis.line = element_blank(),
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        axis.title = element_blank(),
        panel.background = element_blank(),
        panel.border = element_blank(),
        panel.grid = element_blank(),
        plot.title = element_text(hjust = 1.2, size = 20, face = "bold")) +
  ggtitle("Total Sales by Global Publisher")+#setting title
  #Setting the percentage ratio in the center of pie chart
  geom_text(aes(label = paste0(round(total_sales/sum(total_sales) * 100), "%")),
            position = position_stack(vjust = 0.5),color="grey22",size=3)
```


Save a clean version of the chart

```
ggsave("figs/GlobalSales_by_pubs.png",units = "cm",width =20,height = 10,dpi=1000 )
```

**Summary of Visualisation 2**

By analyzing the pie chart, the top ten publishers have obtained over 70% of the share in the video game market. The original data includes over 500 publishers, indicating that although there are many game publishers, most of the sales are actually obtaind by the top publishers, they were significant impact on the overall market, while the other non-top publishers have relatively small impact on the overall market. By observing the development trend of the top ten publishers, we can also roughly predict the overall development of the video game market.

## Summary

After analyzing the data, I have a clearer understanding of the history of video games. However, the visualizations produced so far are still relatively simple and have many limitations. Besides the number of releases, there are many factors that affect global sales, such as the iteration and sale of platforms, which are missing from the original data.

The genre of electronic games is also worth considering. If there is further research, I will classify the genres of electronic games and compare the number of releases and total sales of each genre in the market to understand the preferences of the public. If there is more new data, I also want to focus on studying the tendency of the top ten publishers towards different genres to understand whether the preferences of the public influence the genre choices of the top publishers or whether the top publishers dominate the preferences of the public.