Exploratory Data Analysis - Video Game Sales

Elias Delosreyes

February 19, 2018

Contents

Introduction

The creation of the video game industry opened up a new avenue of entertainment for consumers. Once a small industry, the video game industry has evolved and expanded over the years to become one of the most lucratives sectors of the entertainment business. We will be analyzing the sales and trends of the video game industry, from its infancy in the 1980's to the modern day.

Note: The dataset used contains information on only moderately popular games (100,000+ sales).

Loading and checking the data

First let us load the required R packages needed for the analysis.

```
library(tidyverse) # Used for visuals, reading csv, and manipulating data
library(gridExtra) # Used to create visuals
```

Next we need to load the data into R and check its structure.

```
# Load data
vgsales <- read_csv('vgsales.csv')

# Check data structure
glimpse(vgsales)
```

```
## Observations: 16,598
## Variables: 11
## $ Rank         <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15...
## $ Name         <chr> "Wii Sports", "Super Mario Bros.", "Mario Kart Wi...
## $ Platform     <chr> "Wii", "NES", "Wii", "Wii", "GB", "GB", "DS", "Wi...
## $ Year         <chr> "2006", "1985", "2008", "2009", "1996", "1989", "...
## $ Genre        <chr> "Sports", "Platform", "Racing", "Sports", "Role-P...
## $ Publisher    <chr> "Nintendo", "Nintendo", "Nintendo", "Nintendo", "...
## $ NA_Sales     <dbl> 41.49, 29.08, 15.85, 15.75, 11.27, 23.20, 11.38, ...
## $ EU_Sales     <dbl> 29.02, 3.58, 12.88, 11.01, 8.89, 2.26, 9.23, 9.20...
```

```
## $ JP_Sales     <dbl> 3.77, 6.81, 3.79, 3.28, 10.22, 4.22, 6.50, 2.93, ...
## $ Other_Sales  <dbl> 8.46, 0.77, 3.31, 2.96, 1.00, 0.58, 2.90, 2.85, 2...
## $ Global_Sales <dbl> 82.74, 40.24, 35.82, 33.00, 31.37, 30.26, 30.01, ...
```

We can see there are 16,598 video game titles in this dataset, and 11 different variables that give us some additional insight about each of these.

The variable names are pretty straightforward, but one thing to note is the sales data is not the gross revenue, but the number of copies sold (in terms of millions).

Before we go on, let us alter a few items in the dataset.

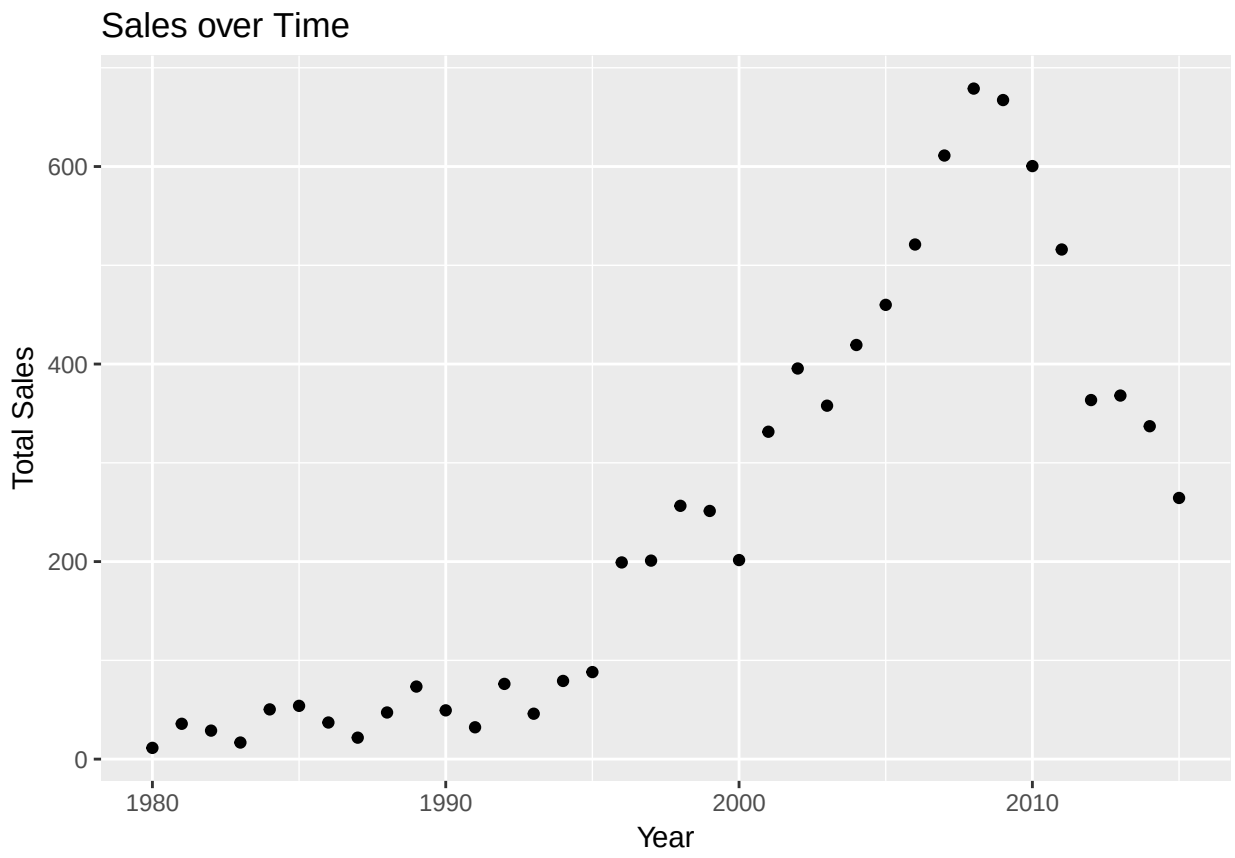
```
# Change data class of `Year`
vgsales$Year <- as.numeric(vgsales$Year)

# List global sales ahead of the other columns
vgsales <- vgsales[,c(1:6, 11, 7:10)]

# We filter out 2016 data, as the dataset was
# collected mid 2016 and the 2016 collection of sales is incomplete.
# We also filter out data with missing Year variable.
vgsales <- vgsales %>%
  filter(Year < 2016, Year != "N/A")
```

Questions

Now that we have seen a glimpse of the data and understand its basic structure. Let's get into asking some questions about this data.

What has the trend of the sales count been over the years?

```
# Group total sales by year
yearly_total_vgsales <- vgsales %>%
  group_by(Year) %>%
  summarize(Total_Sales = sum(Global_Sales, na.rm = FALSE))


# Plot total sales by year
ggplot(yearly_total_vgsales, aes(x = Year, y = Total_Sales)) +
  geom_point()  +
  ggtitle("Sales over Time") +
  ylab('Total Sales')
```



After graphing the data, we can see a clear trend in growth, up until 2010 which is followed by a dramatic decline in sales. This leads us to the next question:

Are sales declining or was the rapid growth unsustainable?

Let's take a look at the top selling games during the period where sales seemed to massively spike.

```
# Find top 10 selling games in 2005 to 2010
vgsales %>%
  filter(Year %in% (2005:2010)) %>%
  arrange(desc(Global_Sales)) %>%
  select(c(1:7)) %>%
  head(10)
```

Rank	Name	Platform	Year	Genre	Publisher	Global_Sales
1	Wii Sports	Wii	2006	Sports	Nintendo	82.74
3	Mario Kart Wii	Wii	2008	Racing	Nintendo	35.82
4	Wii Sports Resort	Wii	2009	Sports	Nintendo	33.00
7	New Super Mario Bros.	DS	2006	Platform	Nintendo	30.01
8	Wii Play	Wii	2006	Misc	Nintendo	29.02
9	New Super Mario Bros. Wii	Wii	2009	Platform	Nintendo	28.62
11	Nintendogs	DS	2005	Simulation	Nintendo	24.76
12	Mario Kart DS	DS	2005	Racing	Nintendo	23.42
14	Wii Fit	Wii	2007	Sports	Nintendo	22.72
15	Wii Fit Plus	Wii	2009	Sports	Nintendo	22.00

```r
# Find the total sales for these 10 games
vgsales%>% filter(Year %in% (2005:2010)) %>%
  arrange(desc(Global_Sales)) %>%
  head(10) %>%
  summarize(sum(Global_Sales), Num_game_titles = n())
```

sum(Global_Sales)	Num_game_titles
332.11	10

```r
# Find the total sales accross all games in this period
vgsales%>% filter(Year %in% (2005:2010)) %>%
  arrange(desc(Global_Sales)) %>%
  summarize(sum(Global_Sales), Num_game_titles = n())
```

sum(Global_Sales)	Num_game_titles
3538.76	7269

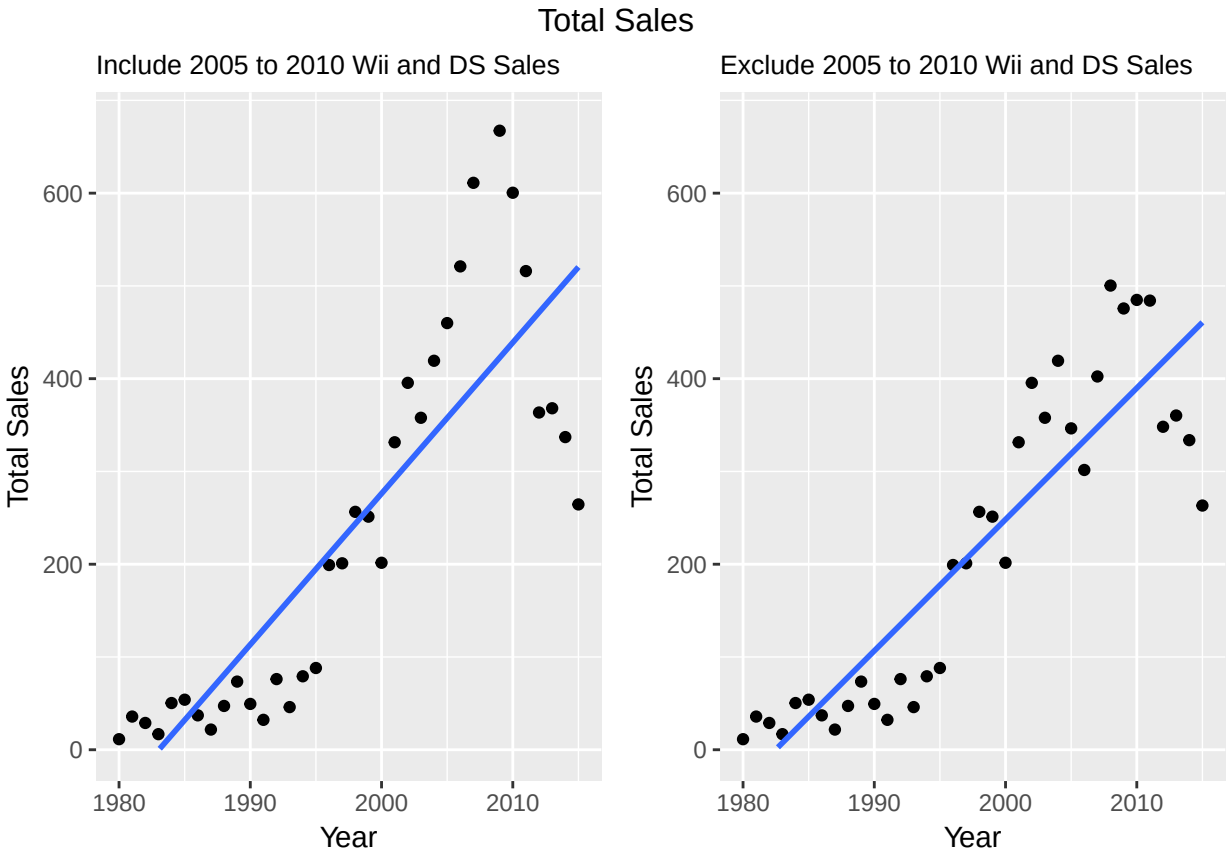
Nintendo seems to dominate the top 10 list of global sales during the years 2005 to 2010. During this time, many of these games were bundled with the Nintendo Wii and Nintendo DS. This context gives some insight into why these Nintendo games seem to dominate the list and why the graph has such an abnormal spike in sales from 2005 to 2010.

In fact, these 10 games account for 332 million sales. Across the 7269 titles produced in this time, theres a total of 3.5 billion sales. In short, the top 0.1% of best selling games account for 10% of the total sales during this period. This top 0.1% is comprised of only Wii and DS games.

With this in mind, lets try and filter out the top selling Wii and DS games during this period. These top selling games may be outliers, and may have temporarily spiked the sales data beyond the current sustainable growth level.

```r
# Filter top 10% games that are Wii or DS from 2005 to 2010
nintendo_outliers <- vgsales %>%
  filter(Platform == 'Wii' | Platform == 'DS' & Year %in% (2005:2010)) %>%
  arrange(desc(Global_Sales)) %>%
  head(300)


# Create dataset that does not include outliers
yearly_total_vgsales2 <- anti_join(vgsales, nintendo_outliers) %>%
  group_by(Year) %>%
  summarize(Total_Sales = sum(Global_Sales, na.rm = FALSE))


# Plot the graphs side-by-side
require(gridExtra)

plot1 <- ggplot(yearly_total_vgsales, aes(x = Year, y = Total_Sales)) +
  geom_point() +
  ylab('Total Sales') +
  ylim(0,675) +
  ggtitle('Include 2005 to 2010 Wii and DS Sales') +
  theme(plot.title = element_text(size = 10)) +
```

```
  geom_smooth(se = FALSE, method = 'lm')


plot2 <- ggplot(yearly_total_vgsales2, aes(x = Year, y = Total_Sales)) +
  geom_point() +
  ylab('Total Sales') + ylim(0,675) +
  ggtitle('Exclude 2005 to 2010 Wii and DS Sales') +
  theme(plot.title = element_text(size = 10)) +
  geom_smooth(se = FALSE,  method = 'lm')


grid.arrange(plot1, plot2, ncol = 2, top = 'Total Sales')
```



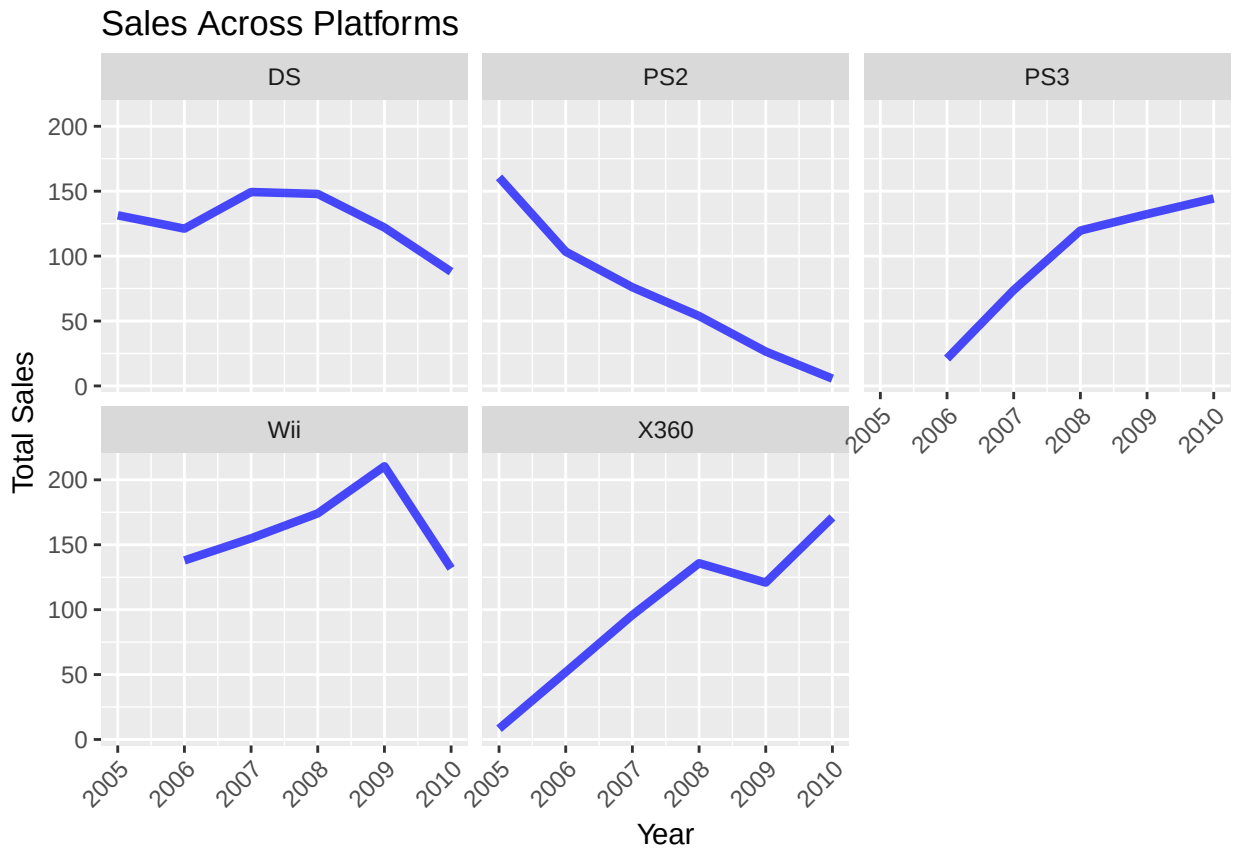
After filtering the top 10% of Wii and DS games, the data points shift, giving a better fit to the trendline. The Nintendo Wii and DS seem to have a heavy pull on global video game sales, and by filtering the top 10% we see a major effect. It could be possible that the massive sales in Wii and DS games spiked the sales beyond sustainable growth levels temporarily. Perhaps we should analyze how other consoles performed during this period to contrast:

What was the growth of sales within the other top consoles during this period?

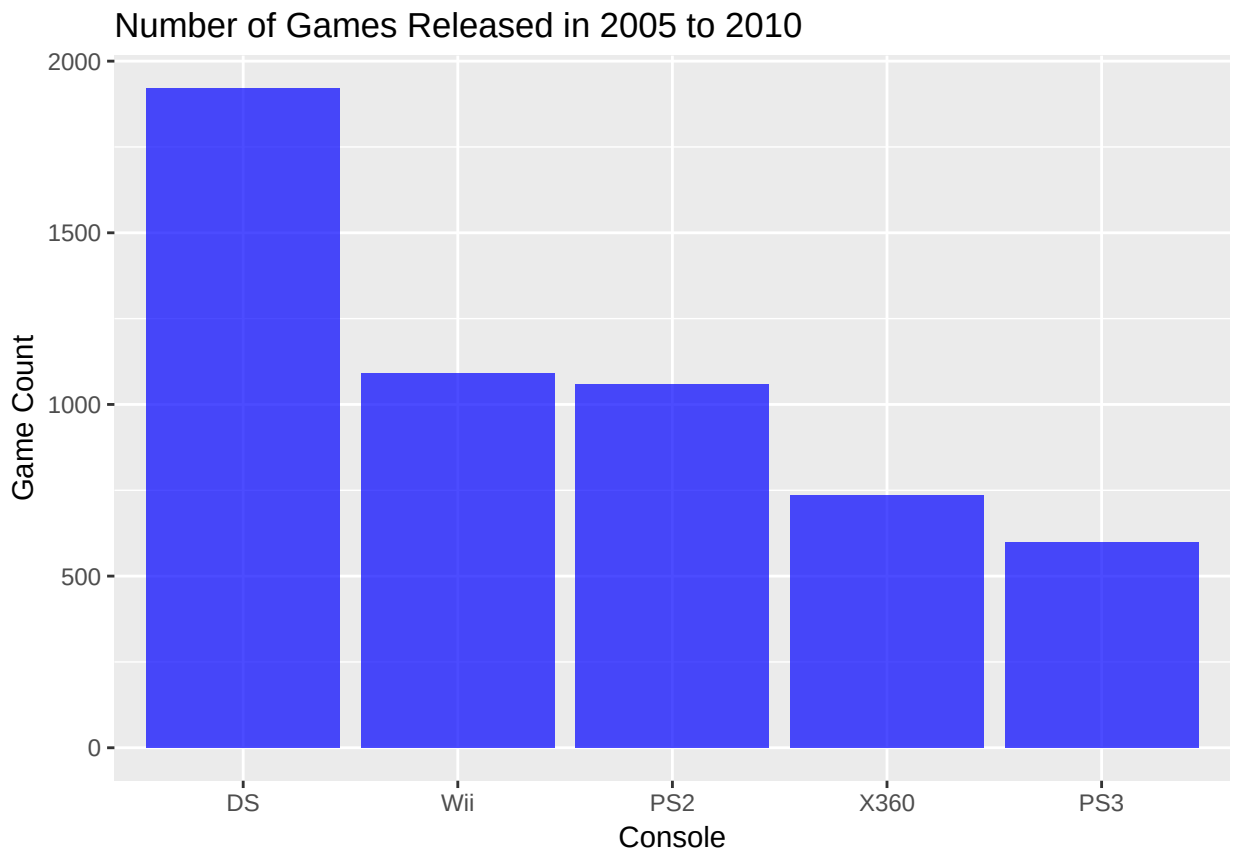
Let's check the growth of the top 5 consoles during 2005 to 2010.

```
# Find total sales for the top 5 consoles for 2005 to 2010
top5_console_2005_2010 <- vgsales %>%
    filter(Year %in% (2005:2010), Platform %in% c('Wii', 'DS', 'PS2', 'PS3', 'X360')) %>%
```

```
    group_by(Year, Platform) %>%
    summarize(Total_Sales = sum(Global_Sales)) %>%
    arrange(Year, desc(Total_Sales))

# Create graph of these sales
ggplot(top5_console_2005_2010, aes(x = Year, y = Total_Sales)) +
  geom_line(size = 1.5, color = 'blue', alpha = 0.7) +
  ylab('Total Sales') +
  ggtitle('Sales Across Platforms') +
  facet_wrap(~ Platform) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



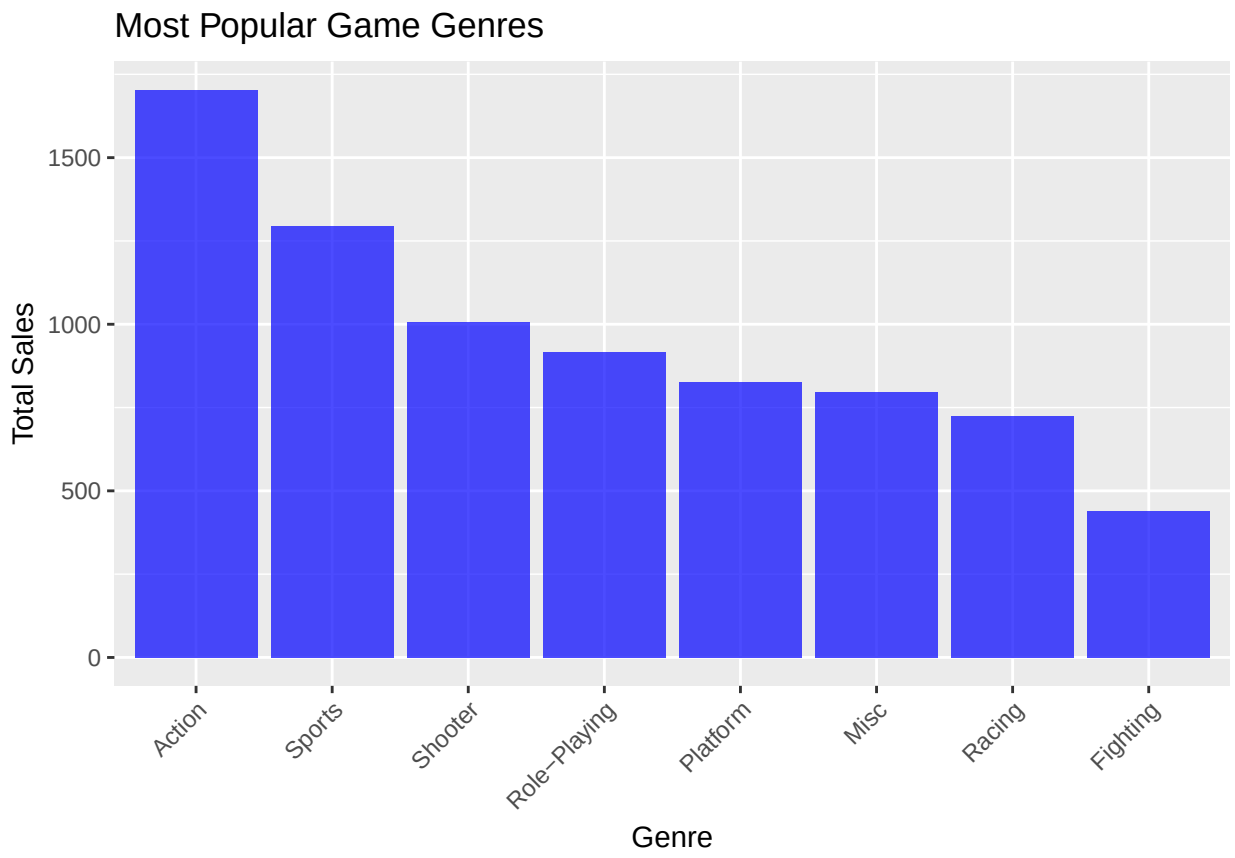

```
# Find total number of games for each console released during this time and graph
vgsales %>%
    filter(Year %in% (2005:2010), Platform %in% c('Wii', 'DS', 'PS2', 'PS3', 'X360')) %>%
    group_by(Platform) %>%
    summarize(Game_count = n()) %>%

ggplot(aes(x = reorder(Platform, -Game_count), y = Game_count))  +
  geom_bar(fill = 'blue', alpha = 0.7, stat = 'identity') +
  xlab('Console') + ylab('Game Count') +
  ggtitle('Number of Games Released in 2005 to 2010')
```



From this we can see the Wii and DS had strong sales but these strong sales may be attributed to them releasing the most games during this period. The PS3 and Xbox 360 had slow sales growth, but performed well for the few games they had released. The PS2 on the other hand, released the third most amount of games, yet still had declining sales, likely due to the PS3 signaling the end of the PS2's lifecycle.

What are the top selling game genres?

```
# Find top genres
top_genres <- vgsales %>%
    group_by(Genre) %>%
    summarize(Total_Sales = sum(Global_Sales)) %>%
    arrange(desc(Total_Sales)) %>%
    head(8)



# Plot top genres
(topgame <- ggplot(top_genres, aes(x = reorder(Genre, -Total_Sales), y = Total_Sales)) +
         geom_bar(fill = 'blue', alpha = 0.7, stat = 'identity') +
          xlab('Genre') + ylab('Total Sales') +
          ggtitle('Most Popular Game Genres') +
          theme(axis.text.x = element_text(angle = 45, hjust = 1)))
```



Action games are the best selling game by far. However there is some possible bias and skew here towards newer games. The gaming industry is constantly expanding, and games in the 2000's naturally have more sales than games before the 2000's, which skews the results.

However, one interesting followup to this for another analysis would be to determine the most popular game genres by decade. For example, in the 1980's maybe action games would not be the most popular, and arcade games with games like Pac-Man or Tetris would show the highest sales.

Summary of Interesting Findings

- **There is a consistent trend in sales growth from the 1980s to the 2000s.**
- **Game sales seem to be declining 2010 and onwards. This may not be the case, but a follow-up with more data would be required**
  - Free games and/or phone games that utilize microtransactions for revenue are not accurately captured in this dataset. As these microtransaction based games have high market value, this is a possible topic for further exploration in another dataset.
  - As the data only includes games with 100,000+ sales, it could be that many recent games (~3 years) are not recorded in the data as they need more time to accrue sales. Thus, this downward trend for game sales in recent years may be deceiving and we would have to revisit this data in the future.
- **From 2005 to 2010, the top 0.1% of of best selling games were for the Nintendo DS and Nintendo Wii.**
  - This top 0.1% accounted for 10% of global sales during this period.
  - During this time, the Nintendo DS released the most game titles by far.
- **Action games are the best selling game genre by far.**