Video Games analysis - demonstration

Data source: Video Game Sales with Ratings

Packages:

```
library(dplyr)
library(ggplot2)
library(tidyr)
library(lubridate)
```

Getting to know the database

Importing the raw database

```
dat <- read.csv("Video_Games_Sales.csv", header=TRUE, na.strings=c("", " ", "NA", "N/A"))
```

```
glimpse(dat)
```

```
## Rows: 16,719
## Columns: 16
## $ Name            <chr> "Wii Sports", "Super Mario Bros.", "Mario Kart Wii"...
## $ Platform        <chr> "Wii", "NES", "Wii", "Wii", "GB", "GB", "DS", "Wii"...
## $ Year_of_Release <int> 2006, 1985, 2008, 2009, 1996, 1989, 2006, 2006, 200...
## $ Genre           <chr> "Sports", "Platform", "Racing", "Sports", "Role-Pla...
## $ Publisher       <chr> "Nintendo", "Nintendo", "Nintendo", "Nintendo", "Ni...
## $ NA_Sales        <dbl> 41.36, 29.08, 15.68, 15.61, 11.27, 23.20, 11.28, 13...
## $ EU_Sales        <dbl> 28.96, 3.58, 12.76, 10.93, 8.89, 2.26, 9.14, 9.18, ...
## $ JP_Sales        <dbl> 3.77, 6.81, 3.79, 3.28, 10.22, 4.22, 6.50, 2.93, 4....
## $ Other_Sales     <dbl> 8.45, 0.77, 3.29, 2.95, 1.00, 0.58, 2.88, 2.84, 2.2...
## $ Global_Sales    <dbl> 82.53, 40.24, 35.52, 32.77, 31.37, 30.26, 29.80, 28...
## $ Critic_Score    <int> 76, NA, 82, 80, NA, NA, 89, 58, 87, NA, NA, 91, NA,...
## $ Critic_Count    <int> 51, NA, 73, 73, NA, NA, 65, 41, 80, NA, NA, 64, NA,...
## $ User_Score      <chr> "8", NA, "8.3", "8", NA, NA, "8.5", "6.6", "8.4", N...
## $ User_Count      <int> 322, NA, 709, 192, NA, NA, 431, 129, 594, NA, NA, 4...
## $ Developer       <chr> "Nintendo", NA, "Nintendo", "Nintendo", NA, NA, "Ni...
## $ Rating          <chr> "E", NA, "E", "E", NA, NA, "E", "E", "E", NA, NA, "...
```

We can see that it contains 16719 games, of which we know 16 parameters.

Most of these variables are clear, NA_Sales means North American Sales and JP_Sales means Japanese Sales. Each of the sales number unit is in millions.

It is also worth noting that all review data is from [Metacritic] (https://www.metacritic.com/) and the Rating column is based on [ESRB] (https://www.esrb.org/about/) content classification standards.

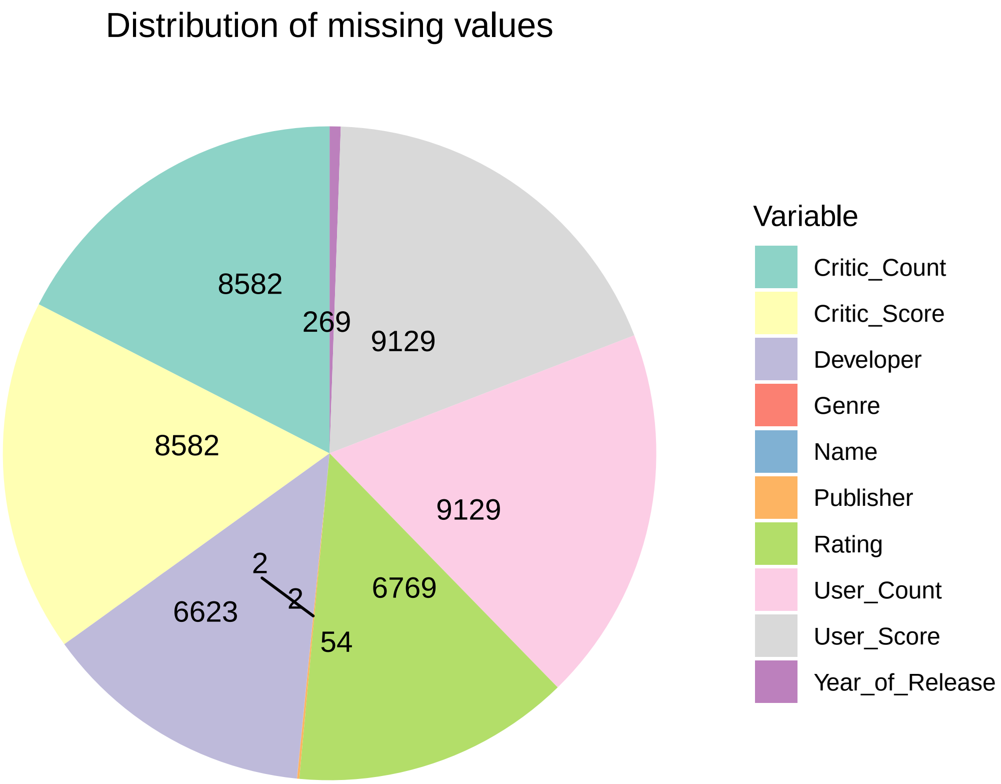
There are two issues that need to be addressed:

```
dat$Year_of_Release <- as.Date(paste(dat$Year_of_Release, 1, 1, sep="-"))
dat$User_Score <- as.numeric(dat$User_Score)
```

The year of release should be converted to date format and the number of user preferences from the incorrect text type to a number for better functionality.

Handling missing values

```
dat %>%
  summarise_all(list(~sum(is.na(.)))) %>%
  gather() %>%
  filter(value!=0) %>%
  ggplot(aes(x=1, y=value, fill=key)) +
  geom_col() +
  ggrepel::geom_text_repel(aes(label=value), position=position_stack(vjust=0.5))+
  coord_polar(theta="y") +
  theme_void()+
  scale_fill_brewer(palette="Set3")+
  labs(fill="Variable")+
  ggtitle("Distribution of missing values")+
  theme(plot.title=element_text(hjust=0.5))
```



Despite the overlap, which is

- Name: 2
- Genre: 2
- Publisher: 54

the missing values revealed themselves. If we are unlucky, it can even cover the entire database. (There is a random game from 2020, but the file was created in 2017)

```
dat <- filter(dat, year(Year_of_Release)<=2017)

dat %>%
  filter(!complete.cases(.)) %>%
  nrow()
```

```
## [1] 9624
```

We are lucky and have “only” 9624 missing rows. Not great, not terrible. We create a new filtered dataset and check for any further missing value.

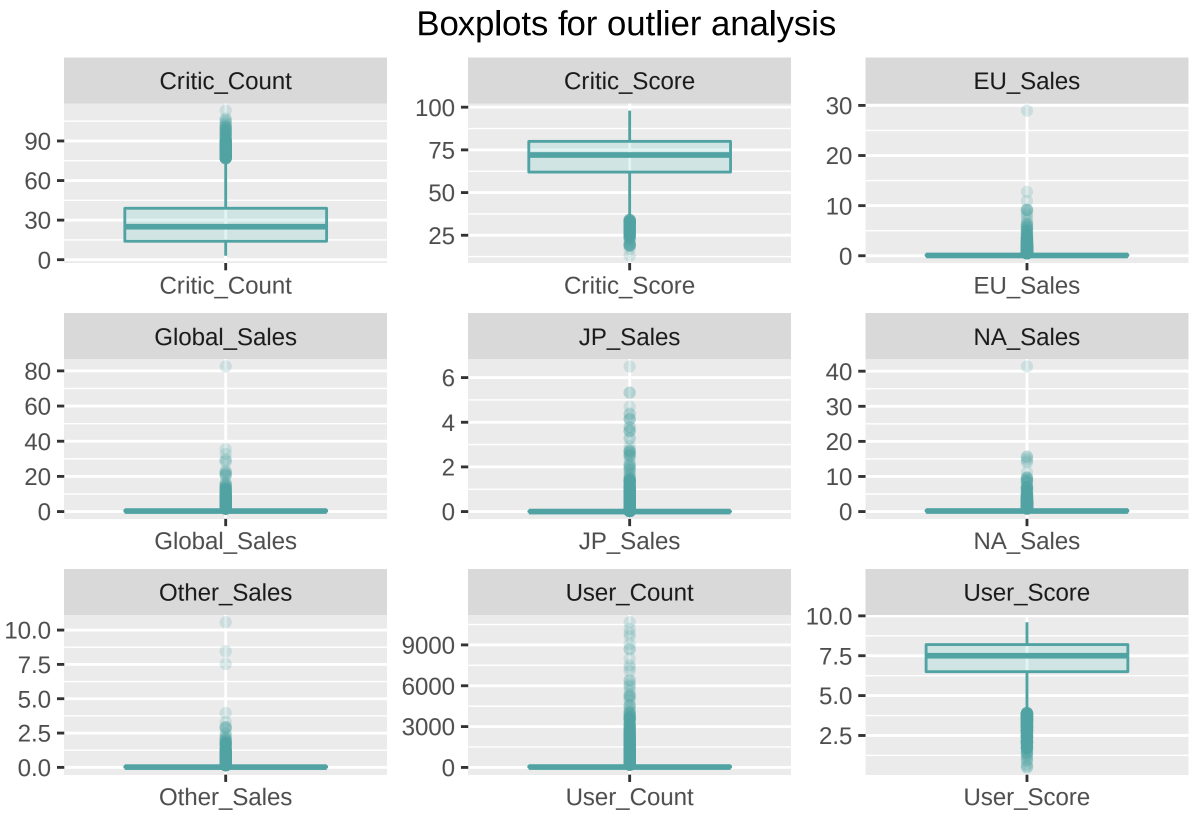
```
no_NA <- dat[complete.cases(dat), ]

sum(!complete.cases(no_NA))
```

```
## [1] 0
```

Outlier analysis

Before starting the analysis we examine the outliers to prevent biases. This is done using a boxplot.

```
no_NA %>%
  select_if(is.numeric) %>%
  gather() %>%
  ggplot(aes(factor(key), value))+
  geom_boxplot(color="#51a3a3", fill="#66CCCC", alpha=0.2)+
  facet_wrap(~key, scale="free")+
  labs(x="", y="")+
  ggtitle("Boxplots for outlier analysis")+
  theme(plot.title=element_text(hjust=0.5))
```



We can see relevant outliers the case of three sales numbers. We don’t know if this is an error or the most interesting factors in our database.

```
cbind(
no_NA %>%
  select(Name, NA_Sales) %>%
  arrange(desc(NA_Sales)) %>%
  top_n(2),

no_NA %>%
  select(Name, EU_Sales) %>%
  arrange(desc(EU_Sales)) %>%
  top_n(2),

no_NA %>%
  select(Name, Global_Sales) %>%
  arrange(desc(Global_Sales)) %>%
  top_n(2))
```

```
## Selecting by NA_Sales
```

```
## Selecting by EU_Sales
```

```
## Selecting by Global_Sales
```

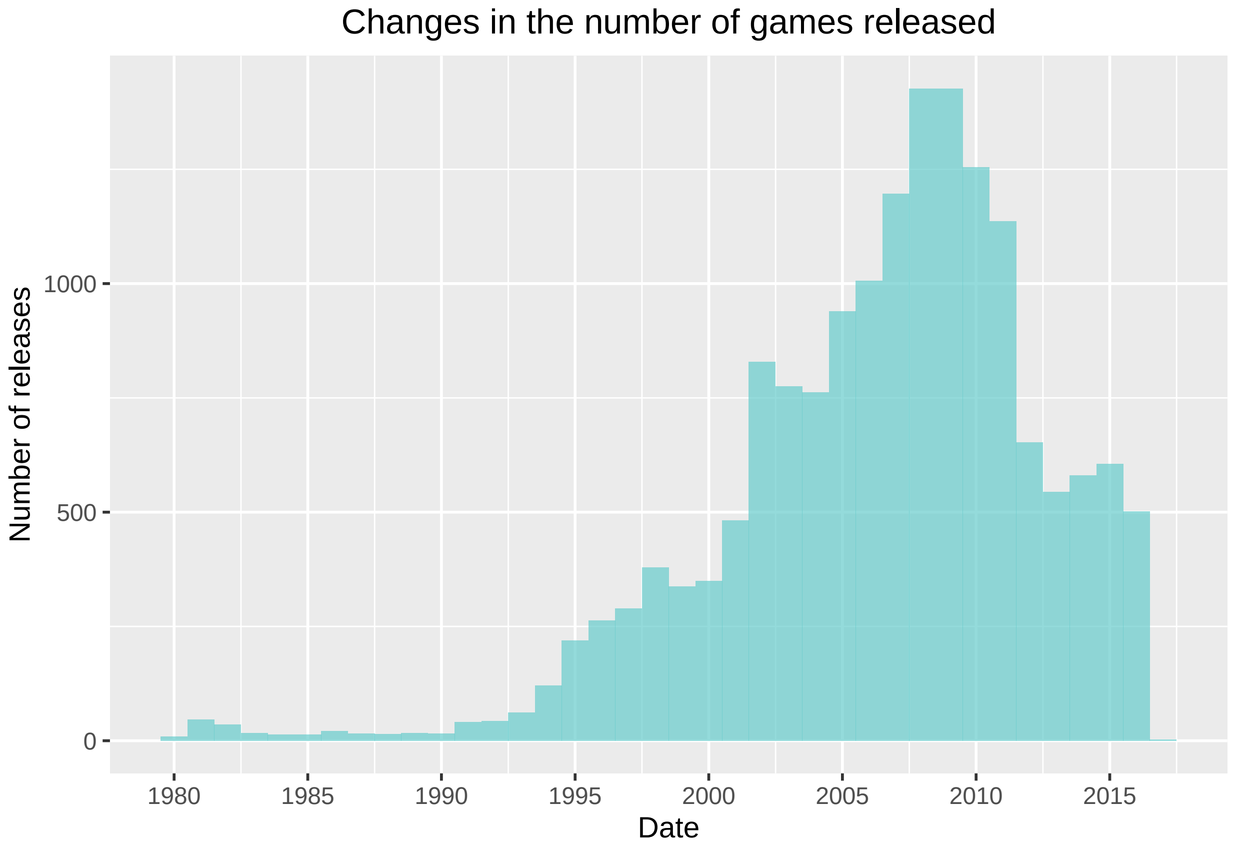
```
##              Name NA_Sales          Name EU_Sales          Name Global_Sales
## 1      Wii Sports    41.36    Wii Sports    28.96    Wii Sports        82.53
## 2 Mario Kart Wii     15.68 Mario Kart Wii    12.76 Mario Kart Wii        35.52
```

If we look at the maximum values for all three variables, we can see that Wii Sports is the outlier. And indeed, the release of this game [was a great success] (https://www.gamespot.com/articles/the-most-influential-games-of-the-21st-century-wii/1100-6466810/).

Exploratory data analysis

First let's look at how the number of games released has evolved over time.

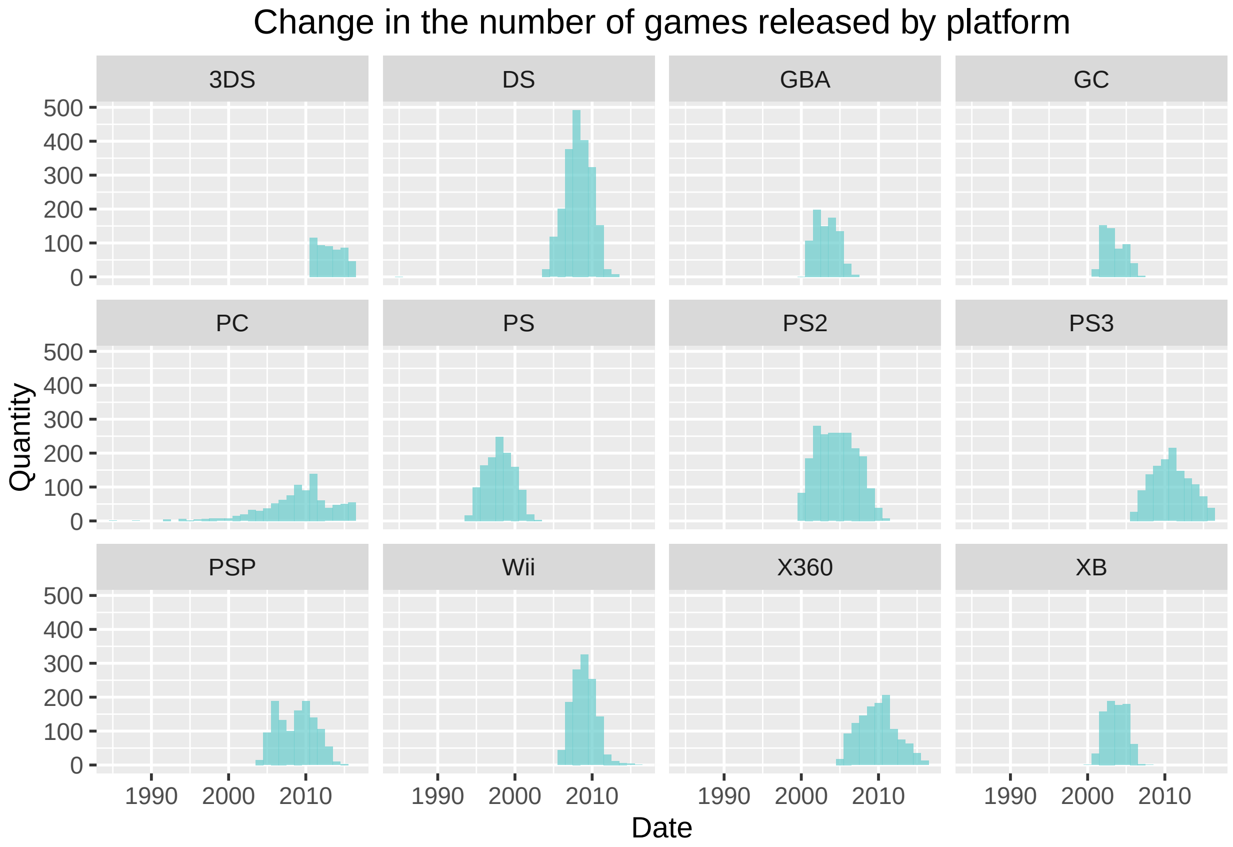
```
ggplot(dat) +
  geom_bar(aes(year(Year_of_Release)), width=1, fill="#66CCCC", alpha=0.7)+
  labs(x="Date", y="Number of releases")+
  ggtitle("Changes in the number of games released")+
  theme(plot.title=element_text(hjust=0.5))+
  scale_x_continuous(breaks = scales::pretty_breaks(10))
```



Until 2009, growth was close to exponential, and then the trend fell even faster to 2002 levels. This is certainly worth exploring further.

Grouping by platforms

```
dat %>%
  group_by(Platform) %>%
  filter(n()>500) %>%
  ggplot() +
    geom_bar(aes(year(Year_of_Release)), width=1, fill="#66CCCC", alpha=0.7)+
    labs(x="Date", y="Quantity")+
    ggtitle("Change in the number of games released by platform")+
    scale_fill_brewer(palette="Set3")+
    theme(plot.title=element_text(hjust=0.5))+
    facet_wrap(.~Platform)
```

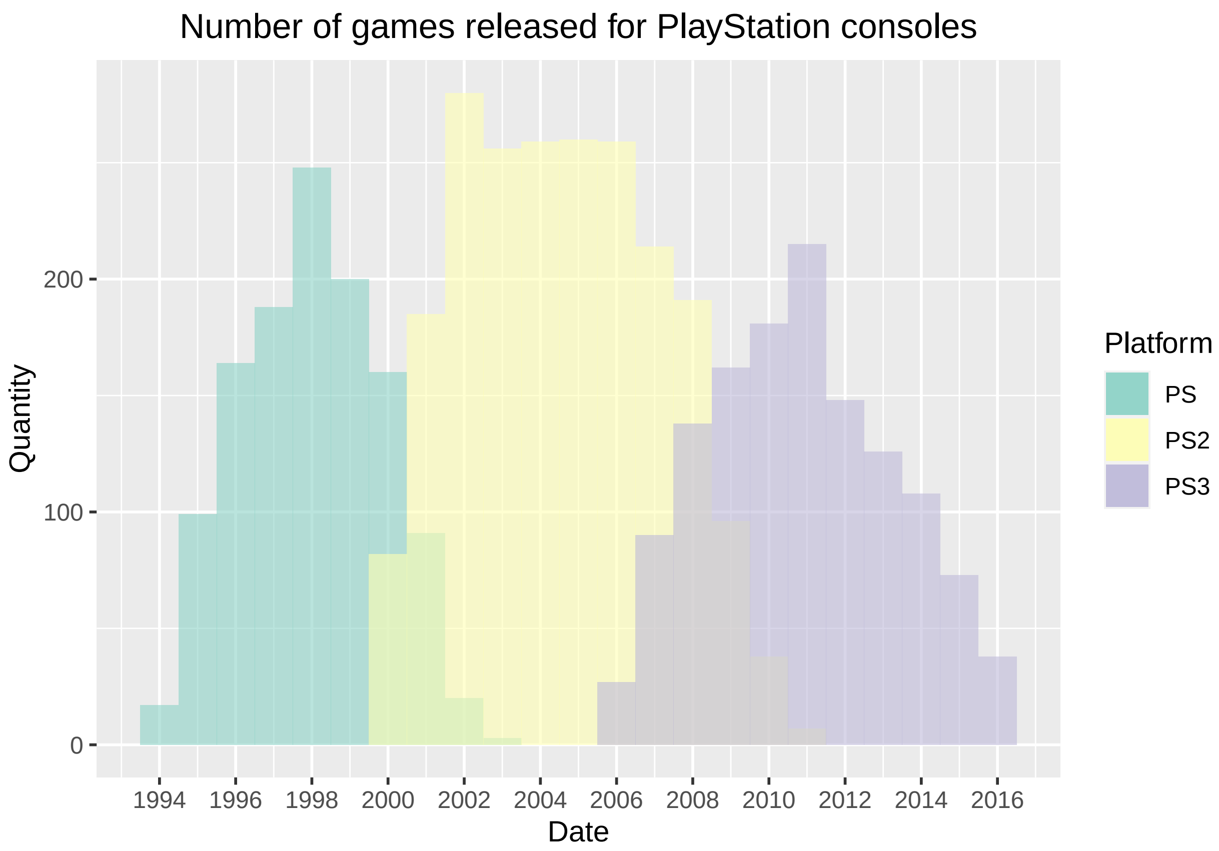


If we look at the game release numbers of the "most active" platforms, we can see that this is largely due to the Nintendo DS and Nintendo Wii. A few days after the DS, the PSP also appeared (similar style), which also contributed to the 2009 peak.

What is strange that there was also a decline in the release of PC games after the peak in 2009, which means that the change in the trend cannot be explained only by the emergence and decay of new platforms.

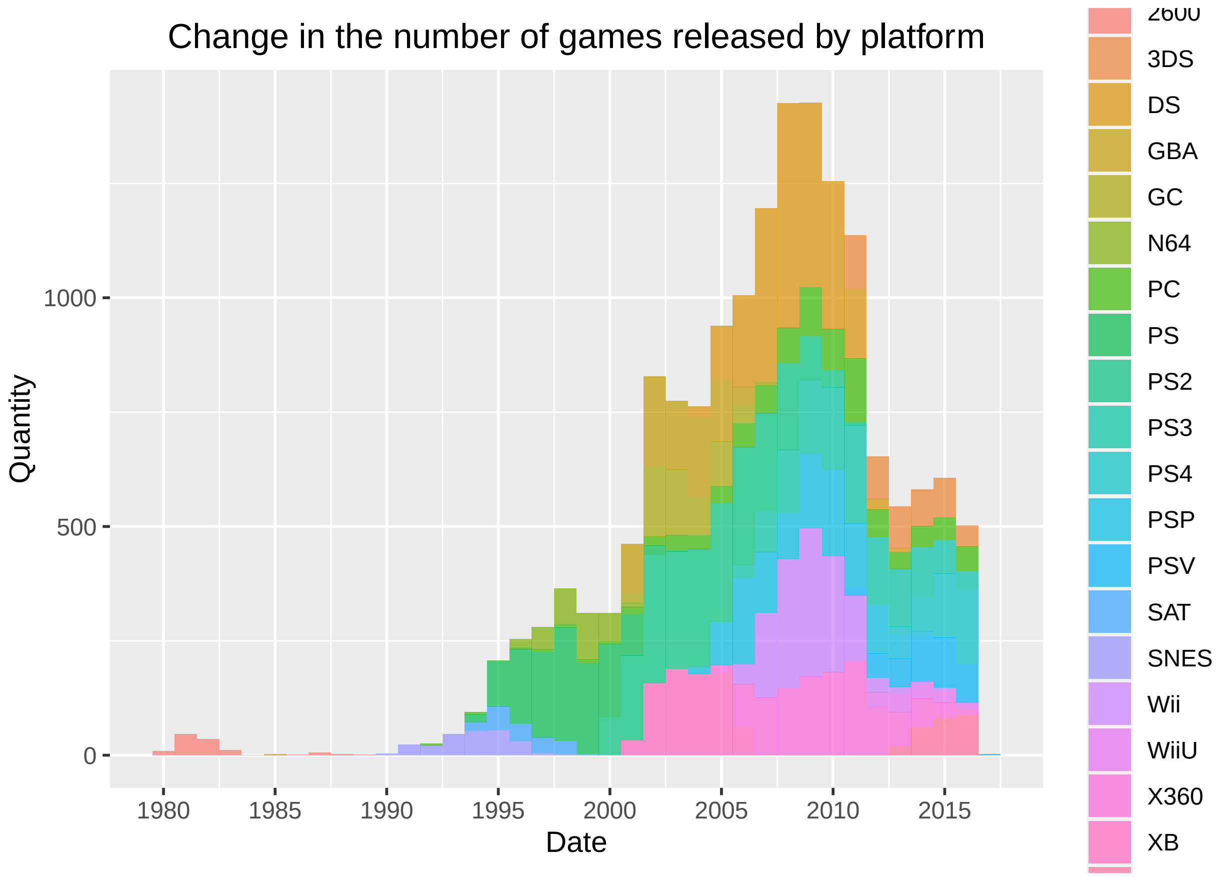
For the sake of curiosity, we can also see the lifespan of PlayStation consoles:

```r
ggplot()+
  geom_bar(data=filter(dat, Platform=="PS"), aes(year(Year_of_Release), fill=Platform), width=1, alpha=
  geom_bar(data=filter(dat, Platform=="PS2"), aes(year(Year_of_Release), fill=Platform), width=1, alpha
  geom_bar(data=filter(dat, Platform=="PS3"), aes(year(Year_of_Release), fill=Platform), width=1, alpha
  labs(x="Date", y="Quantity")+
  ggtitle("Number of games released for PlayStation consoles")+
  scale_fill_brewer(palette="Set3")+
  theme(plot.title=element_text(hjust=0.5))+
  scale_x_continuous(breaks = scales::pretty_breaks(10))
```



Let's also look at different platforms.

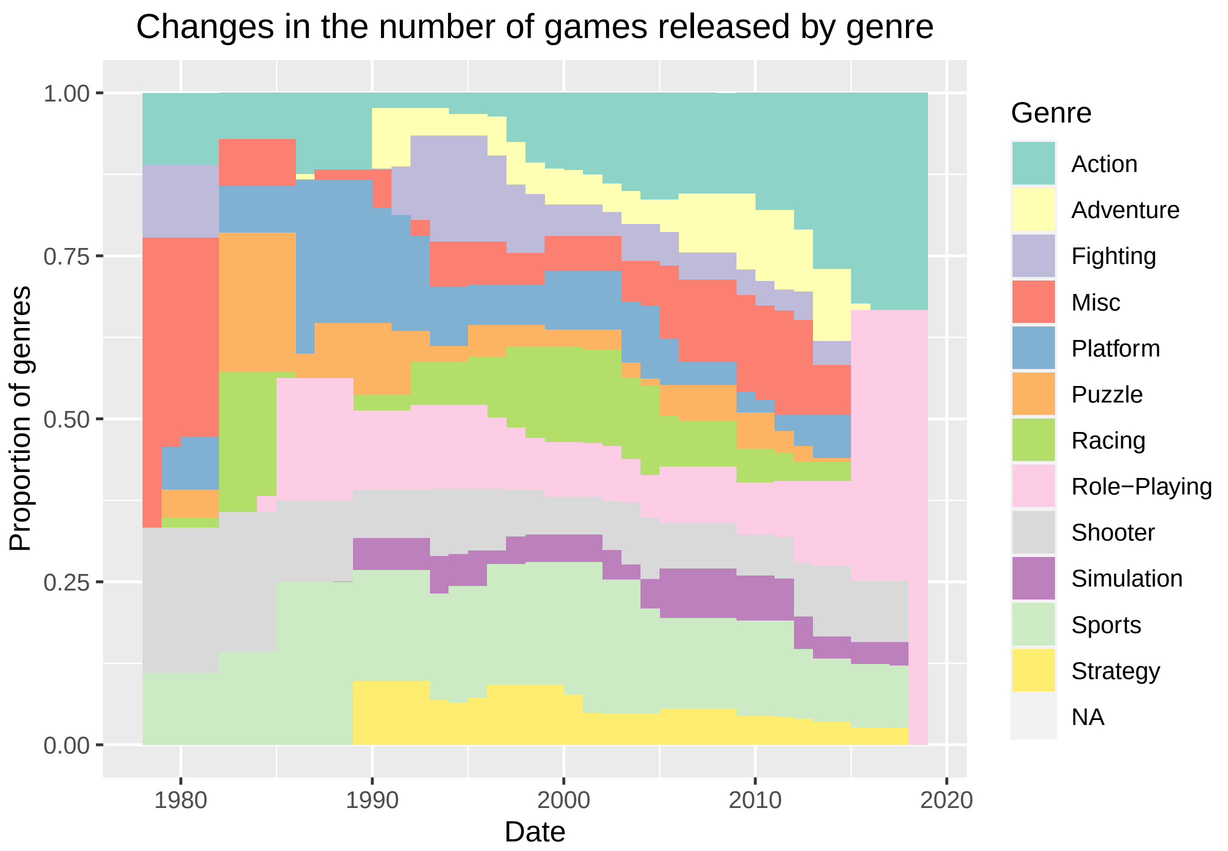
```
dat %>%
  group_by(Platform) %>%
  filter(n()>100) %>%
  ggplot(aes(fill=Platform)) +
  geom_bar(mapping = aes(year(Year_of_Release)), width=1, alpha=0.7)+
  labs(x="Date", y="Quantity")+
  ggtitle("Change in the number of games released by platform")+
  theme(plot.title=element_text(hjust=0.5))+
  scale_x_continuous(breaks = scales::pretty_breaks(10))
```



We can see that there was a large selection in the market during the peak period. Perhaps we can say that the golden age of arcade games in the 80s was repeated?

Grouping by genre

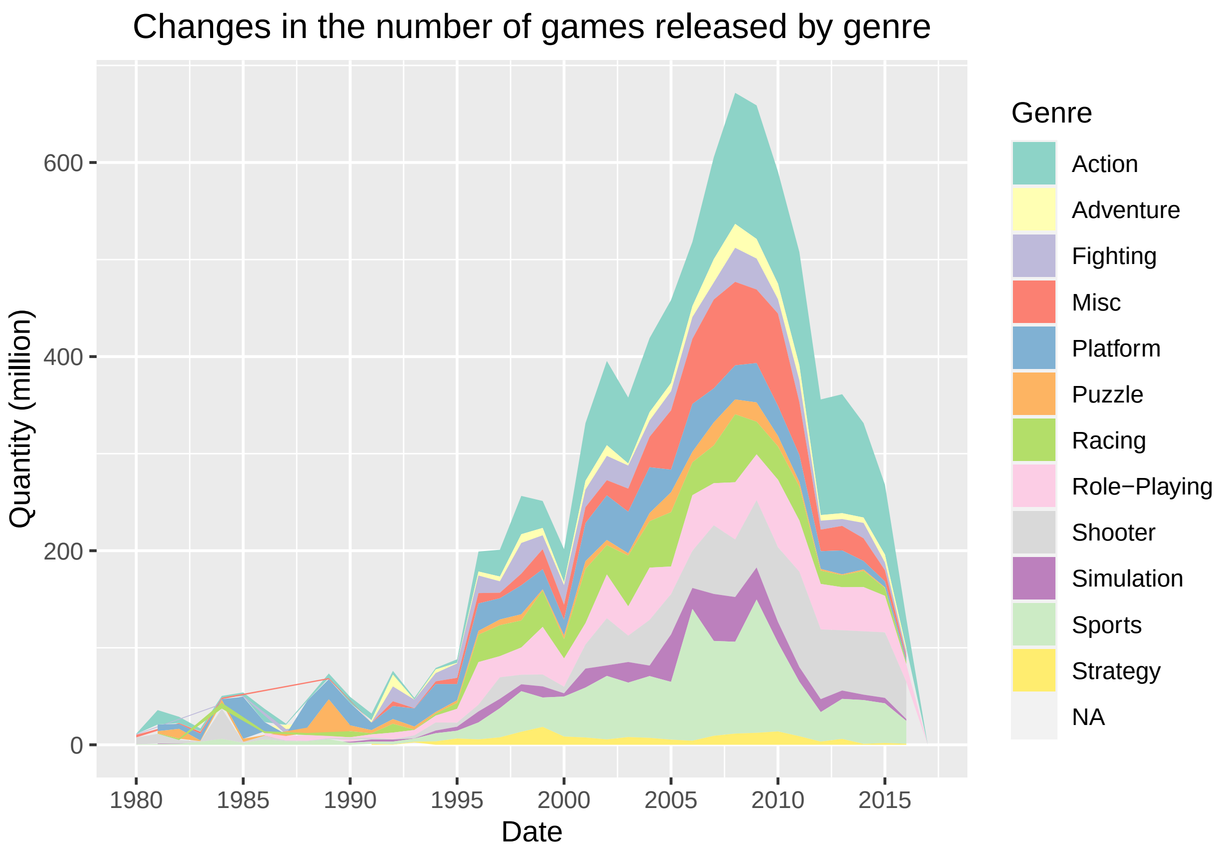
```
dat %>%
  ggplot(aes(fill=Genre)) +
  geom_bar(position="fill", mapping = aes(year(Year_of_Release)), width=4)+
  labs(x="Date", y="Proportion of genres")+
  scale_fill_brewer(palette="Set3")+
  ggtitle("Changes in the number of games released by genre")+
  theme(plot.title=element_text(hjust=0.5))
```



On this 100% stacked bar chart, we can follow the development and appearance of the number of games of each genre.

What we can see is that action games are starting to dominate the market. The proportion of representatives of most genres is really stagnant, strategy, sports, racing games are declining, while role-playing games have started to grow again.

```
dat %>%
  group_by(Year_of_Release, Genre) %>%
  summarise(Global_Sales=sum(Global_Sales)) %>%
  ggplot(aes(Year_of_Release, Global_Sales, fill=Genre))+
  geom_area()+
  xlab("Date")+
  ylab("Quantity (million)")+
  ggtitle("Changes in the number of games released by genre")+
  theme(plot.title=element_text(hjust=0.5))+
  scale_fill_brewer(palette="Set3")+
  scale_x_date(breaks = scales::pretty_breaks(10))
```



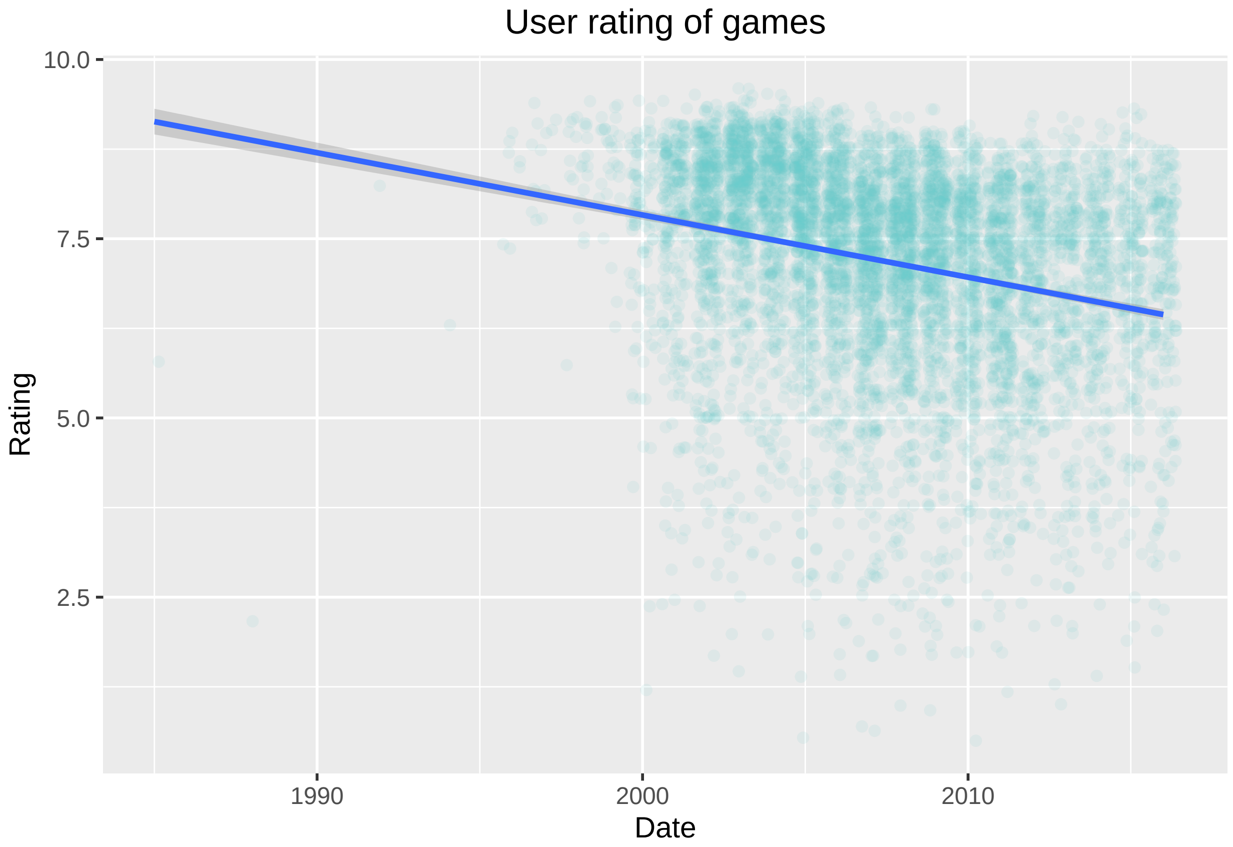
To examine not only the ratios but also their extent, the area diagram above may be useful.

Analysis of reviews

Let's look at how Metacritic user ratings have evolved over the years and fit a linear regression trend to it.

```
no_NA %>%
  ggplot(aes(Year_of_Release, User_Score))+
  geom_jitter(alpha=0.1, color="#66CCCC")+
  geom_smooth(method="lm")+
  labs(x="Date", y="Rating")+
  ggtitle("User rating of games")+
  theme(plot.title=element_text(hjust=0.5))
```

```
## `geom_smooth()` using formula 'y ~ x'
```

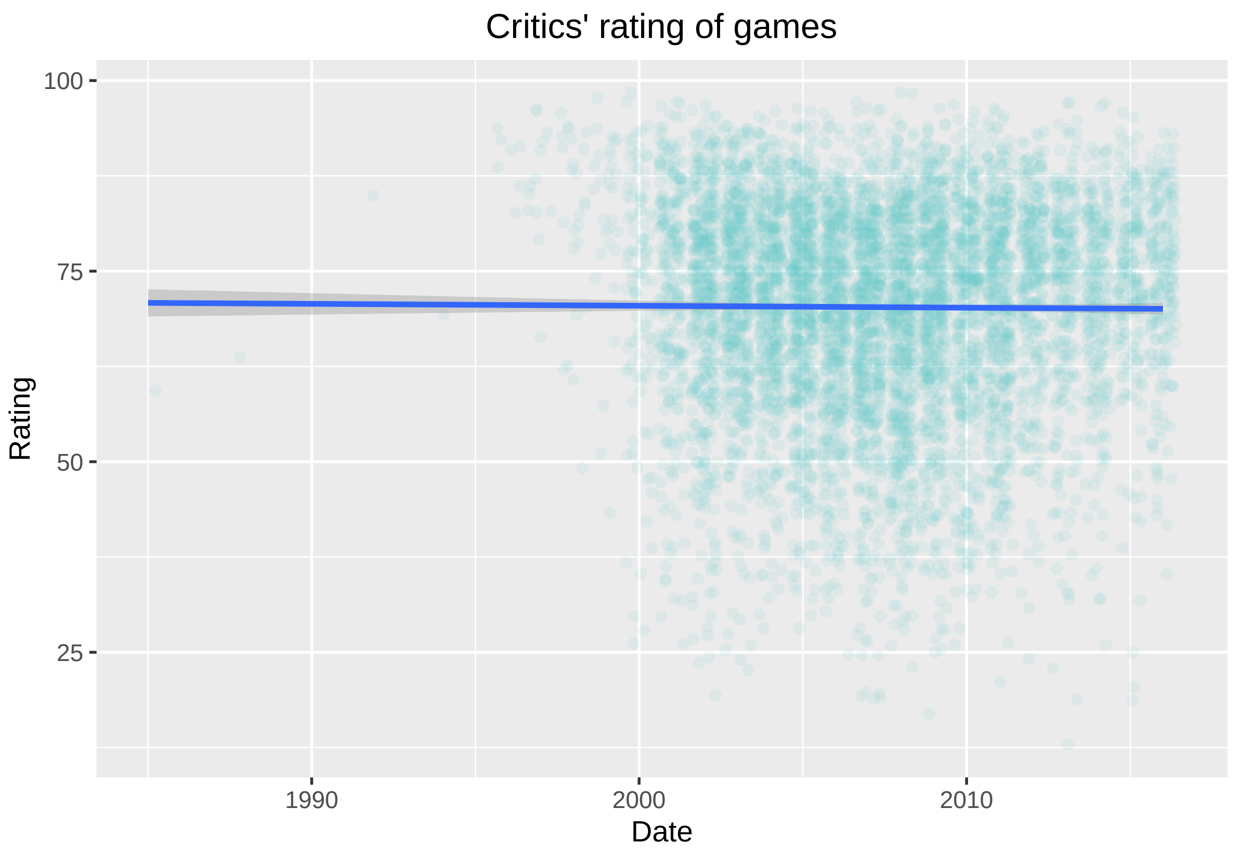


Based on both the scatterplot and the linear trend, there appears to be a decline in player satisfaction. Of course, this may also be related to the increase and dilution of the number of games.

Let's look at the same chart based on the reviews of critics.

```
no_NA %>%
  ggplot(aes(Year_of_Release, Critic_Score))+
  geom_jitter(alpha=0.1, color="#66CCCC")+
  geom_smooth(method="lm")+
  labs(x="Date", y="Rating")+
  ggtitle("Critics' rating of games")+
  theme(plot.title=element_text(hjust=0.5))
```

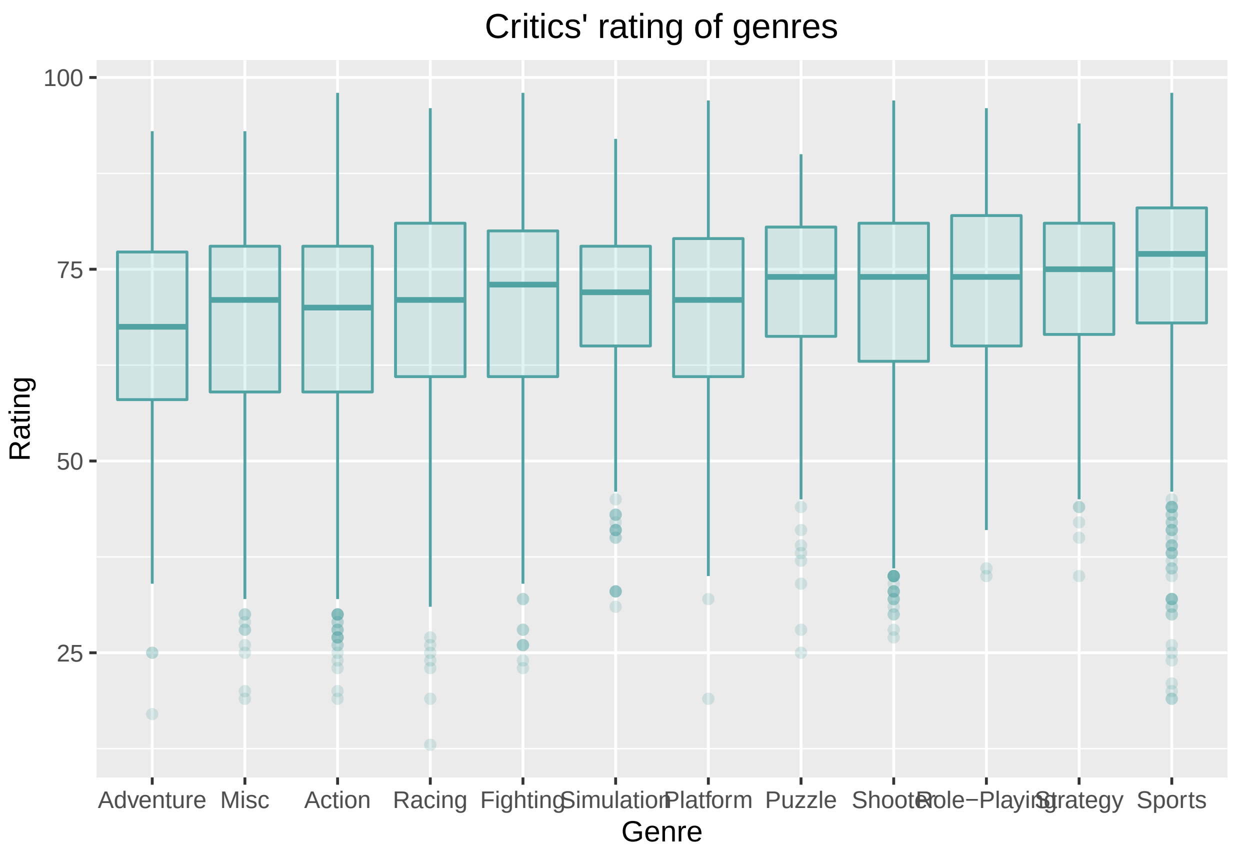
```
## `geom_smooth()` using formula 'y ~ x'
```



Interestingly, the average rating of critical opinions has not changed significantly over the years.

I'll take a little closer look at this:

```r
no_NA %>%
  ggplot(aes(reorder(Genre, Critic_Score, function(x) + mean(x)), Critic_Score))+
  geom_boxplot(color="#51a3a3", fill="#66CCCC", alpha=0.2)+
  labs(x="Genre", y="Rating")+
  ggtitle("Critics' rating of genres")+
  theme(plot.title=element_text(hjust=0.5))
```

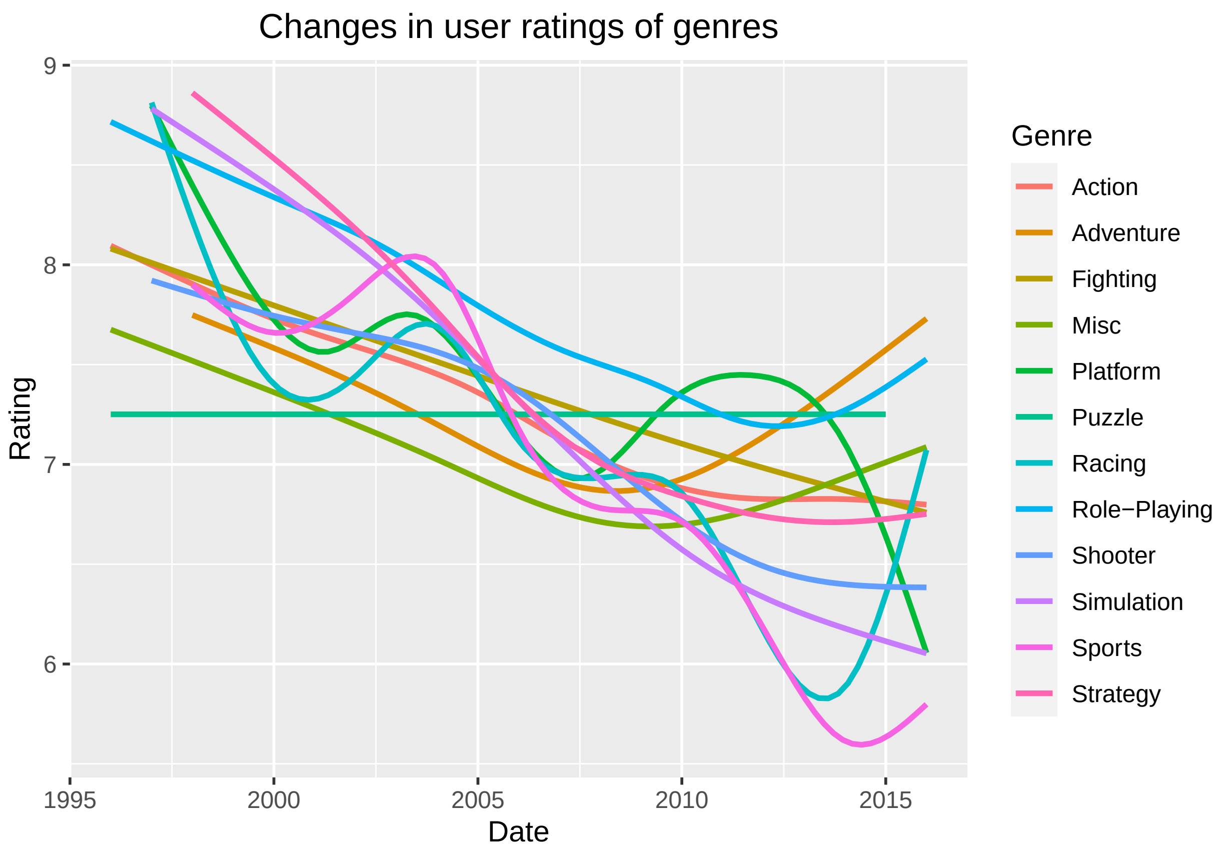


We can see that there are plenty of downward outliers in all categories, and the average rating of action games is certainly one of the lowest, while sports and strategy games have the best average rating.

Check how gamers rated it in contrast from year to year

```
no_NA %>%
  filter(year(Year_of_Release)>1995) %>%
  ggplot(aes(Year_of_Release, User_Score, color=Genre))+
  geom_smooth(se=FALSE)+
  labs(x="Date", y="Rating")+
  ggtitle("Changes in user ratings of genres")+
  theme(plot.title=element_text(hjust=0.5))
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

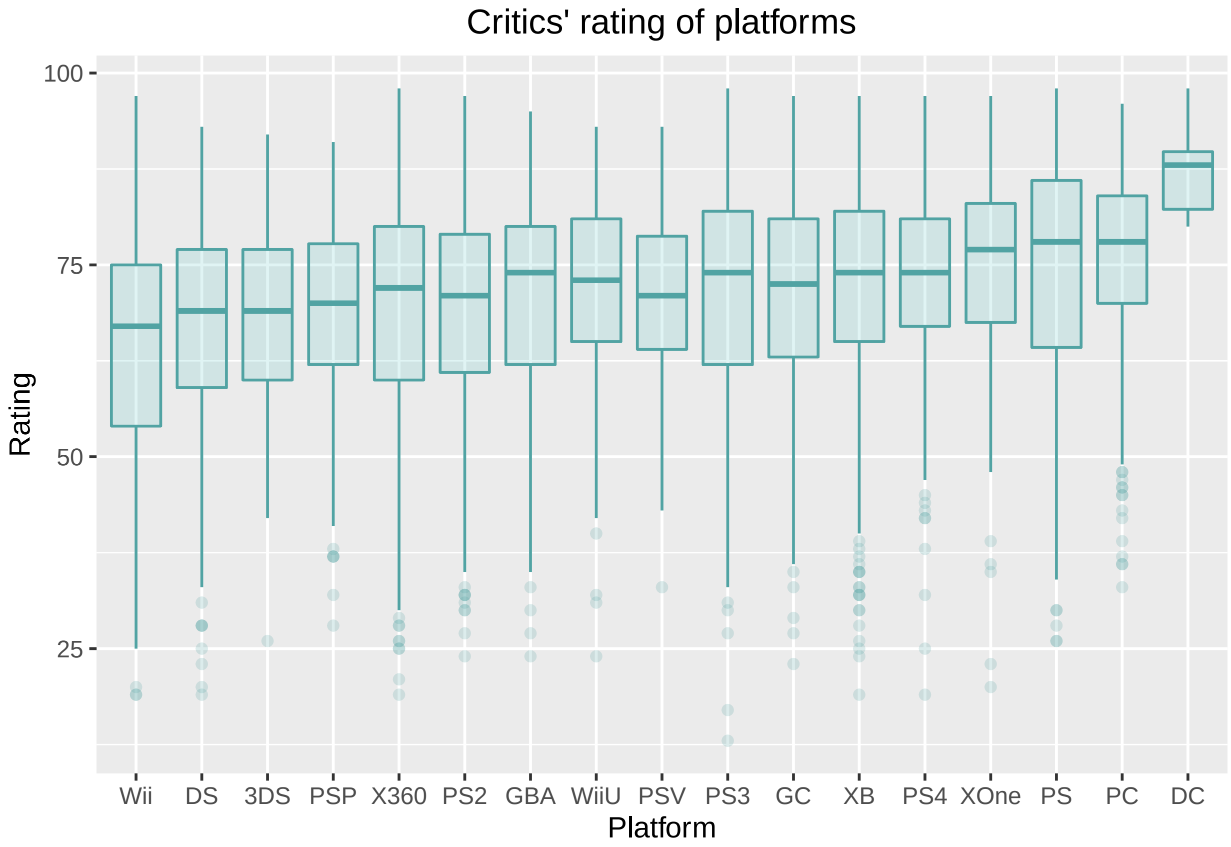


The rating of most genres has worsened over time, the rating of adventure games and role-playing games has improved in recent years.

The ratings for strategy, platform, and logic games are quite variable, while those for sports and simulation games have worsened in the most significant way.

Now that we know how critics and gamers think about genres, let's look at their reviews on platforms.

```
no_NA %>%
  ggplot(aes(reorder(Platform, Critic_Score, function(x) + mean(x)), Critic_Score))+
  geom_boxplot(color="#51a3a3", fill="#66CCCC", alpha=0.2)+
  labs(x="Platform", y="Rating")+
  ggtitle("Critics' rating of platforms")+
  theme(plot.title=element_text(hjust=0.5))
```

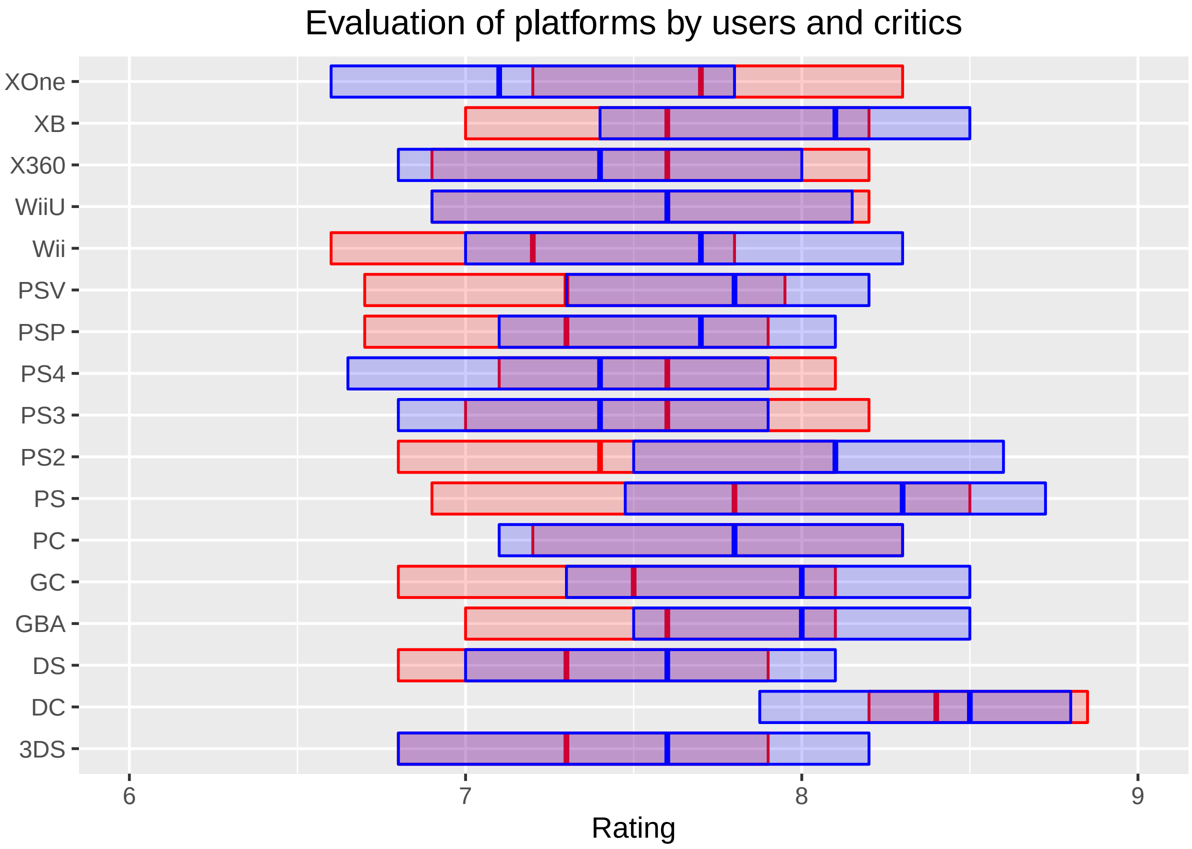


Although Wii games have been sold in large numbers, critics still consider them the platform with the worst games. clearly the games on the Dreamcast platform got the best rating, this is the Sega 1998 console anyway, and it looked like this (interestingly, the controller also had a small screen):

Of course, the fact that there are far fewer games than other platforms contributes to the good rating, so for example, the PC with the second best games is perhaps an even bigger win.

It may be interesting to see how the opinions of critics and users meet on each platform.

```
no_NA %>%
  ggplot()+
  geom_boxplot(aes(Platform, Critic_Score/10, ymin=..lower.., ymax=..upper..), fill="red", color="red",
  geom_boxplot(aes(Platform, User_Score, ymin=..lower.., ymax=..upper..), fill="blue", color="blue", al
  labs(x="", y="Rating")+
  ggtitle("Evaluation of platforms by users and critics")+
  theme(plot.title=element_text(hjust=0.5))+
  coord_flip()+
  ylim(6,9)
```



In red we see the interquartile scales of the critics and in blue the interquartile volumes of the users' evaluations as well as their average. Surprisingly, they rarely agree.

Users are much less satisfied with games on Xbox One, while critics are more dissatisfied with Xbox, Wii, PS Vita, PS2, PS and GameCube platforms

TOP games

There is nothing left but to examine the real giants, with the highest number of games sold each year (which is part of the data set). As the sign of respect, here is the full list:

```
dat %>%
  group_by(Name) %>%
  group_by(Year_of_Release) %>%
  top_n(1, Global_Sales) %>%
  arrange(Year_of_Release) %>%
  mutate(year = year(Year_of_Release)) %>%
  ungroup() %>%
  select(year, Name, Global_Sales) %>%
  knitr::kable("html", col.names=c("Year", "Game", "International sales (million units)"), align="c")
```

Year

Game

International sales (million units)

1980

Asteroids

4.31

1981

Pitfall!

4.50

1982

Pac-Man

7.81

1983

Baseball

3.20

1984

Duck Hunt

28.31

1985

Super Mario Bros.

40.24

1986

The Legend of Zelda

6.51

1987

Zelda II: The Adventure of Link

4.38

1988

Super Mario Bros. 3

17.28

1989

Tetris

30.26

1990

Super Mario World

20.61

1991

The Legend of Zelda: A Link to the Past

4.61

1992

Super Mario Land 2: 6 Golden Coins

11.18

1993

Super Mario All-Stars

10.55

1994

Donkey Kong Country

9.30

1995

Donkey Kong Country 2: Diddy's Kong Quest

5.15

1996

Pokemon Red/Pokemon Blue

31.37

1997

Gran Turismo

10.95

1998

PokĂ©mon Yellow: Special Pikachu Edition

14.64

1999

Pokemon Gold/Pokemon Silver

23.10

2000

PokĂ©mon Crystal Version

6.39

2001

Gran Turismo 3: A-Spec

14.98

2002

Grand Theft Auto: Vice City

16.15

2003

Need for Speed Underground

7.20

2004

Grand Theft Auto: San Andreas

20.81

2005

Nintendogs

24.67

2006

Wii Sports

82.53

2007

Wii Fit

22.70

2008

Mario Kart Wii

35.52

2009

Wii Sports Resort

32.77

2010

Kinect Adventures!

21.81

2011

Call of Duty: Modern Warfare 3

14.73

2012

Call of Duty: Black Ops II

13.79

2013

Grand Theft Auto V

21.04

2014

Grand Theft Auto V

12.61

2015

Call of Duty: Black Ops 3

14.63

2016

FIFA 17

7.59

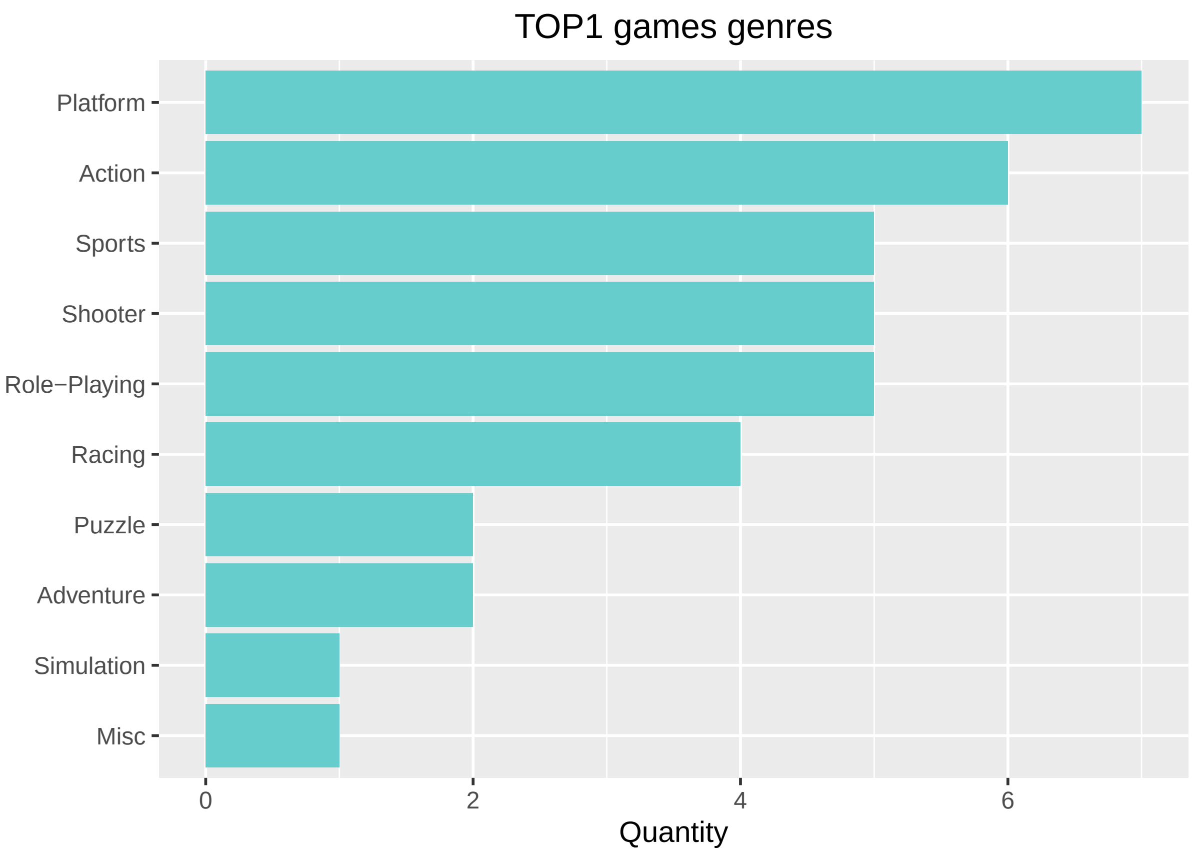
2017

Phantasy Star Online 2 Episode 4: Deluxe Package

0.04

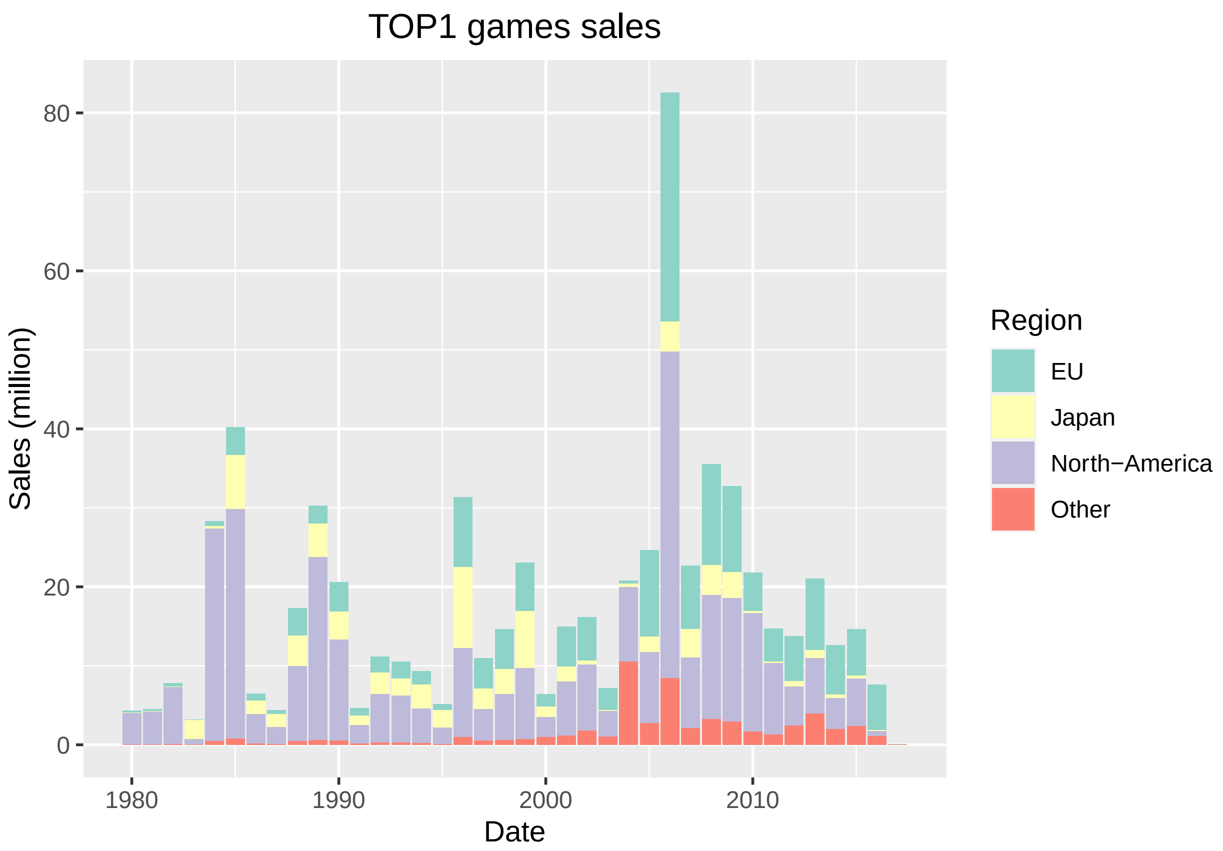
We may be interested in what genre most giants came from.

```r
dat %>%
  group_by(Name) %>%
  group_by(Year_of_Release) %>%
  top_n(1, Global_Sales) %>%
  ggplot(aes(reorder(Genre,Genre,function(x)+length(x))))+
  geom_bar(fill="#66CCCC")+
  coord_flip()+
  labs(x="", y="Quantity")+
  ggtitle("TOP1 games genres")+
  theme(plot.title=element_text(hjust=0.5))
```



By platform games we mean those usually retro games where you have to jump from platforms, i.e. platforms, and run forward (backwards) in a 2D environment.

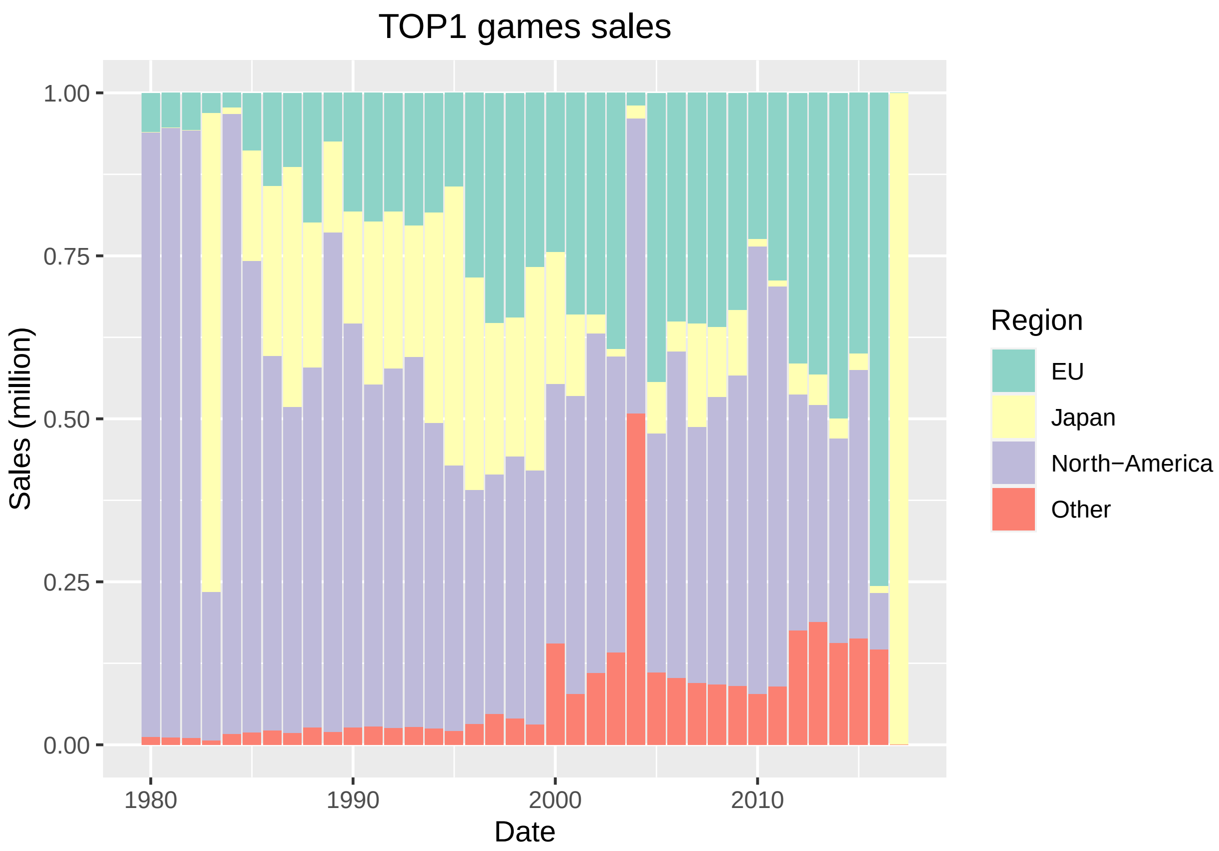
```
dat %>%
  group_by(Name) %>%
  group_by(Year_of_Release) %>%
  top_n(1, Global_Sales) %>%
  select(Year_of_Release, NA_Sales, EU_Sales, JP_Sales, Other_Sales) %>%
  gather(type, count, -Year_of_Release) %>%
  ggplot(aes(as.Date(Year_of_Release, "%Y"), count, fill=factor(type)))+
  geom_bar(stat="identity")+
  labs(x="Date", y="Sales (million)", fill="Region")+
  ggtitle("TOP1 games sales")+
  theme(plot.title=element_text(hjust=0.5))+
  scale_fill_brewer(palette="Set3", labels=c("EU", "Japan", "North-America", "Other"))
```



The first chart shows that before the 2000s, the average TOP1 game sales were relatively low, while there are also games with particularly high numbers.

After 2000, this chart shape changed and a much more predictable chart emerged.

```
dat %>%
  group_by(Name) %>%
  group_by(Year_of_Release) %>%
  top_n(1, Global_Sales) %>%
  select(Year_of_Release, NA_Sales, EU_Sales, JP_Sales, Other_Sales) %>%
  gather(type, count, -Year_of_Release) %>%
  ggplot(aes(as.Date(Year_of_Release, "%Y"), count, fill=factor(type)))+
  geom_bar(stat="identity", position="fill")+
  labs(x="Date", y="Sales (million)", fill="Region")+
  ggtitle("TOP1 games sales")+
  theme(plot.title=element_text(hjust=0.5))+
  scale_fill_brewer(palette="Set3", labels=c("EU", "Japan", "North-America", "Other"))
```



The second 100% stacked figure shows well that Europe's buying power has become increasingly significant, while Japan's has almost disappeared. North America is showing relative stability, and other regions have become significant since the turn of the millennium.

I think we analysed every aspect of the database and we were able to draw a lot of interesting conclusions.