Professional Certificate in Data Science

Videogames Project

Elaborated by: Marco Méndez Atienza

July 21, 2021

Contents

Required packages

```
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")
if(!require(data.table)) install.packages("data.table", repos = "http://cran.us.r-project.org")
if(!require(ggthemes)) install.packages("ggthemes", repos = "http://cran.us.r-project.org")
if(!require(ggrepel)) install.packages("ggrepel", repos = "http://cran.us.r-project.org")
if(!require(viridis)) install.packages("viridis", repos = "http://cran.us.r-project.org")
if(!require(RColorBrewer)) install.packages("RColorBrewer", repos = "http://cran.us.r-project.org")
if(!require(lubridate)) install.packages("lubridate", repos = "http://cran.us.r-project.org")
if(!require(ggmosaic)) install.packages("ggmosaic", repos = "http://cran.us.r-project.org")
if(!require(readr)) install.packages("readr", repos = "http://cran.us.r-project.org")
if(!require(stargazer)) install.packages("stargazer", repos = "http://cran.us.r-project.org")

library(tidyverse)
library(caret)
library(data.table)
library(ggthemes)
library(ggrepel)
library(viridis)
library(RColorBrewer)
library(lubridate)
library(ggmosaic)
library(stargazer)
```

1 Introduction

The video games are a billionaire industry that has been around for many years. Since the first games were published around the 1970s, they have come a long way to have hyper realistic graphics and huge entire worlds inside them.

According to Statista ("Video Game Industry - Statistics & Facts," 2021), the revenue for this industry in 2020 was estimated to be almost 37 billion USD.Compared to previous years, now the generation of game consumers has a significant larger spending power; contrary to what was believed, games are no longer exclusive for children, they top the preferences of many people across generations,

To give some insights about the current context of game industry, consider the following facts:

- The most successful public companies in terms of gaming revenue are Tencent, Sony and Apple.
- PlayStation 4 was the best seller of the past generation consoles.
- The favorite genres in 2020 were action, strategy and adventure.
- The most popular store for hard copy games was GameStop, while Amazon was the most popular for digital purchases.
- It is estimated that around 18% of adults spend 6 or more hours per week playing video games.

The succinct introduction is sufficient to justify a statistical and econometric analysis for the available data of this industry. For the objectives of this report, a public data set was used; specifically, one provided by the website Kaggle, which overall description can be found **here**. Effectively, this data set contains information about video game sales and ratings (given by users and critics) from 1980 to 2017.

It is sensible to assume that games rating directly affect their commercial success. Nonetheless, it is possible to obtain much deeper insights regarding other variables.

Thus, the following pages answer the following questions:

1. How does different types of game ratings (users and critics) correlate between each other?

The hypothesis for this question is that both ratings behave similarly and do not display major dissonances between each other.

2. Does the commercial success of games is positively correlated with the ratings given by users and critics?

For this question, the hypothesis is that, in general, ratings positively and clearly explain the units sold of games.

The rest of the report is divided in three parts:

1. Exploratory Data Analysis: meaningful and useful descriptive statistics are presented to gain insights about the distributions and tendencies of video game sales.
2. Linear Models: simple linear models are fitted to display the relationships between different facets of ratings and sales, which provide interesting and surprising results in several cases.
3. Conclusion: the most relevant and pertinent results are summarized and compared to the hypothesis raised.

2 Exploratory Data Analysis

The data set can be found in the following path:

```
library(readr)

videogames <- read_csv("videogames.csv")
```

```
head(videogames)
```

```
## # A tibble: 6 x 16
##   Name      Platform Year_of_Release Genre  Publisher NA_Sales EU_Sales JP_Sales
##   <chr>     <chr>    <chr>           <chr>  <chr>        <dbl>    <dbl>    <dbl>
## 1 Wii Spor~ Wii      2006            Sports Nintendo      41.4    29.0      3.77
## 2 Super Ma~ NES      1985            Platf~ Nintendo      29.1     3.58     6.81
## 3 Mario Ka~ Wii      2008            Racing Nintendo      15.7    12.8      3.79
## 4 Wii Spor~ Wii      2009            Sports Nintendo      15.6    10.9      3.28
## 5 Pokemon ~ GB       1996            Role-~ Nintendo      11.3     8.89    10.2
## 6 Tetris    GB       1989            Puzzle Nintendo      23.2     2.26     4.22
## # ... with 8 more variables: Other_Sales <dbl>, Global_Sales <dbl>,
## #   Critic_Score <dbl>, Critic_Count <dbl>, User_Score <chr>, User_Count <dbl>,
## #   Developer <chr>, Rating <chr>
```

As we can see, some string values are "N/A," so lets replace them:

```
videogames[videogames == "N/A"] <- NA
```

First of all, lets describe the variables the data set includes:

```
names(videogames)
```

```
##  [1] "Name"            "Platform"        "Year_of_Release" "Genre"
##  [5] "Publisher"       "NA_Sales"        "EU_Sales"        "JP_Sales"
##  [9] "Other_Sales"     "Global_Sales"    "Critic_Score"    "Critic_Count"
## [13] "User_Score"      "User_Count"      "Developer"       "Rating"
```

- "Name": Name of the game
- "Platform": Console on which the game is running
- "Year_of_Release": Year of the game released
- "Genre": Game's category
- "Publisher": Publisher
- "NA_Sales": Games sales in North America (in millions of units)
- "EU_Sales": Games sales in the European Union (in millions of units)
- "JP_Sales": Games sales in Japan (in millions of units)
- "Other_sales": Games sales in the rest of the world, i.e. Africa, Asia excluding Japan, Australia, Europa excluding the EU, and Latin America (in millions of units)
- "Global_sales": Total sales in the world (in millions of units)
- "Critic_Score": Aggregate score compiled by Metacritic staff
- "Critic_Count": The number of critics used in coming up with the Critic_Score
- "User_Score": Score by Metacritic's subscribers
- "User_Count": Number of users who gave the User_Score

- "Developer": Party responsible for creating the game}
- "Rating": The ESRB rating (e.g. Everyone, Teens, Adults, etc.)

Additionally, this data set contains information for 16 719 observations and 16 different variables:

```
dim(videogames)
```

```
## [1] 16719    16
```

Lets perform some exploratory data analysis for each variable in order to obtain some meaningful insights about the data. In the following pages, the games in the data set will be referred as "published games," regardless of not including some other known games. Also, consider that one relevant variable is year: this data set covers information from 1980 to 2017, effectively.

First, notice that some games were released in multiple platforms (e.g. 007: Quantum of Solace was released in PC, DS, PS2, Wii, PS3 and XBox 360); thus, it is useful to know how many unique games the data set includes:

```
length(unique(videogames$Name))
```

```
## [1] 11563
```

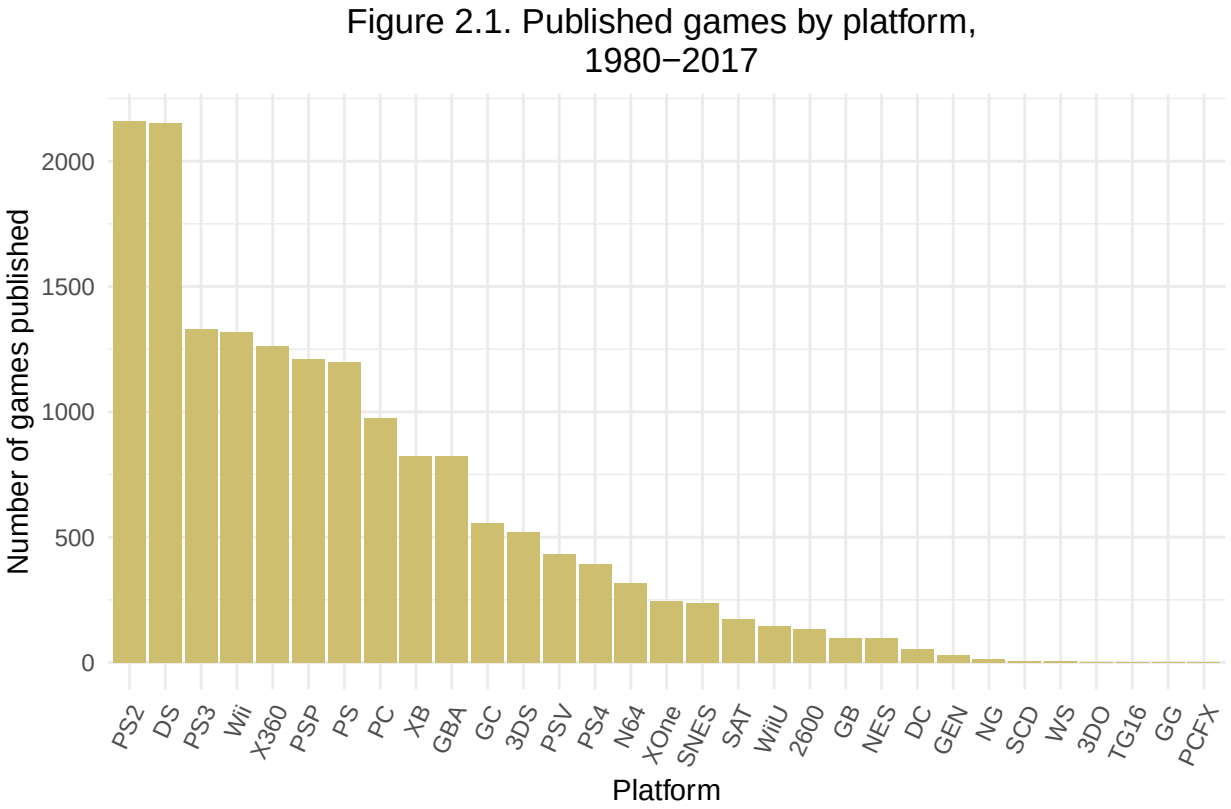
Lets now focus on the platform, genre and publisher: PS2 and DS are the platforms with the most published games (more than 2 000) within the aforementioned time interval, while NES, DC, GEN, NG, SCD, WS, 3DO, TG18, GG and PCFX have all less than 100 published games.

```
table(videogames$Platform)
```

```
##
## 2600  3DO  3DS   DC   DS   GB  GBA   GC  GEN   GG  N64  NES   NG   PC PCFX   PS
##  133    3  520   52 2152   98  822  556   29    1  319   98   12  974    1 1197
##  PS2  PS3  PS4  PSP  PSV  SAT  SCD SNES TG16  Wii WiiU   WS X360   XB XOne
## 2161 1331  393 1209  432  173    6  239    2 1320  147    6 1262  824  247
```

The following graph shows number of games released by platform, in decreasing order:

```
videogames %>%
  ggplot(aes(x = reorder(Platform, Platform, function(x) - length(x)))) +
  geom_bar(fill = "lightgoldenrod3") +
  labs(title = "Figure 2.1. Published games by platform, \n 1980-2017",
       y = "Number of games published",
       x = "Platform",
       caption = "Source: Own elaboration based on the data 'videogames' from kaggle ") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text.x = element_text(angle = 65, hjust = 1), plot.caption = element_text(hjust= 0))
```



Source: Own elaboration based on the data 'videogames' from kaggle

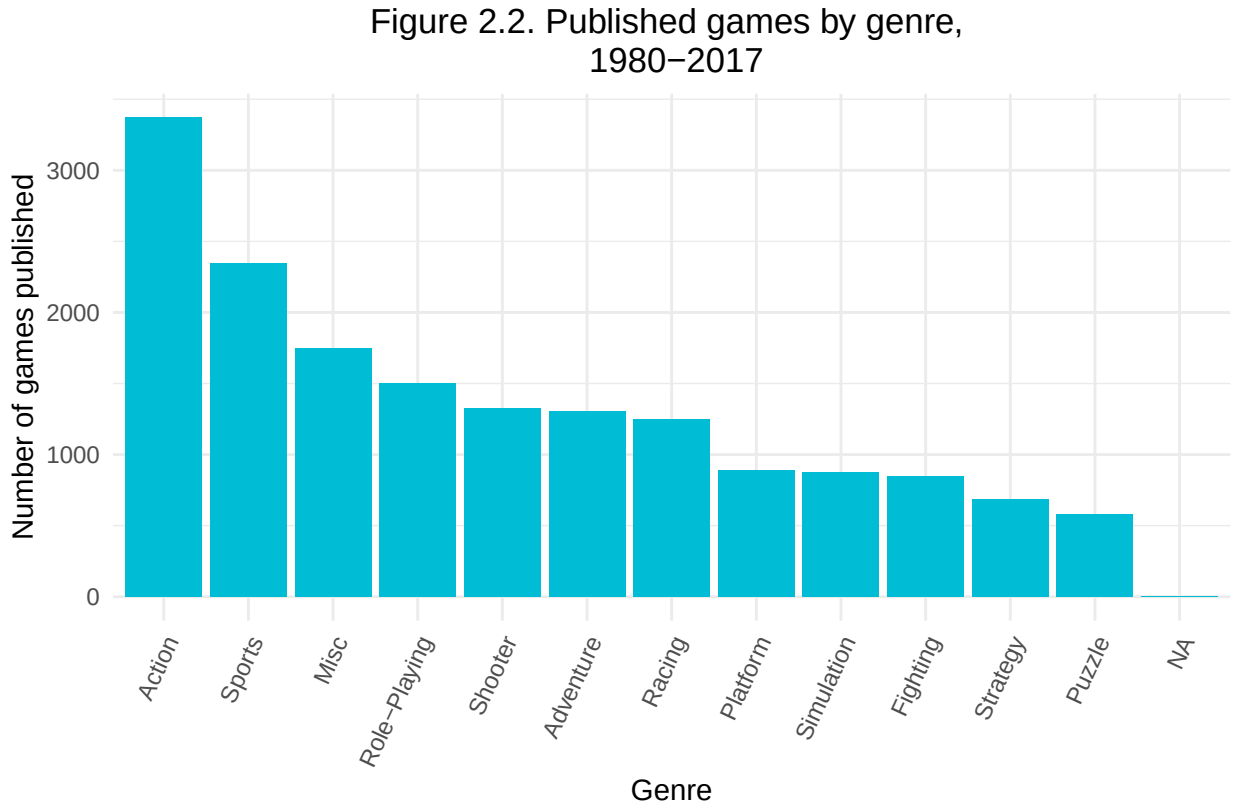
Consider now the game genre. "Action" appears to be, by far, the most common genre in videogames, followed by sports and Misc, while Fighting, Strategy and Puzzle are the least commons.

```
table(videogames$Genre)
```

```
##
##       Action    Adventure     Fighting         Misc     Platform       Puzzle
##         3370         1303          849         1750          888          580
##       Racing Role-Playing      Shooter   Simulation       Sports     Strategy
##         1249         1500         1323          874         2348          683
```

We can graphically see that in the follwing plot:

```
videogames %>%
  ggplot(aes(x = reorder(Genre, Genre, function(x) - length(x)))) +
  geom_bar(fill = "#00BCD4") +
  labs(title = "Figure 2.2. Published games by genre, \n 1980-2017",
       y = "Number of games published",
       x = "Genre",
       caption = "Source: Own elaboration based on the data 'videogames' from kaggle ") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text.x = element_text(angle = 65, hjust = 1), plot.caption = element_text(hjust= 0))
```



Source: Own elaboration based on the data 'videogames' from kaggle

Another pertinent insight is to look at the number of games published by publisher. Firstly, there are 581 different publishers in the data set, a rather big figure.

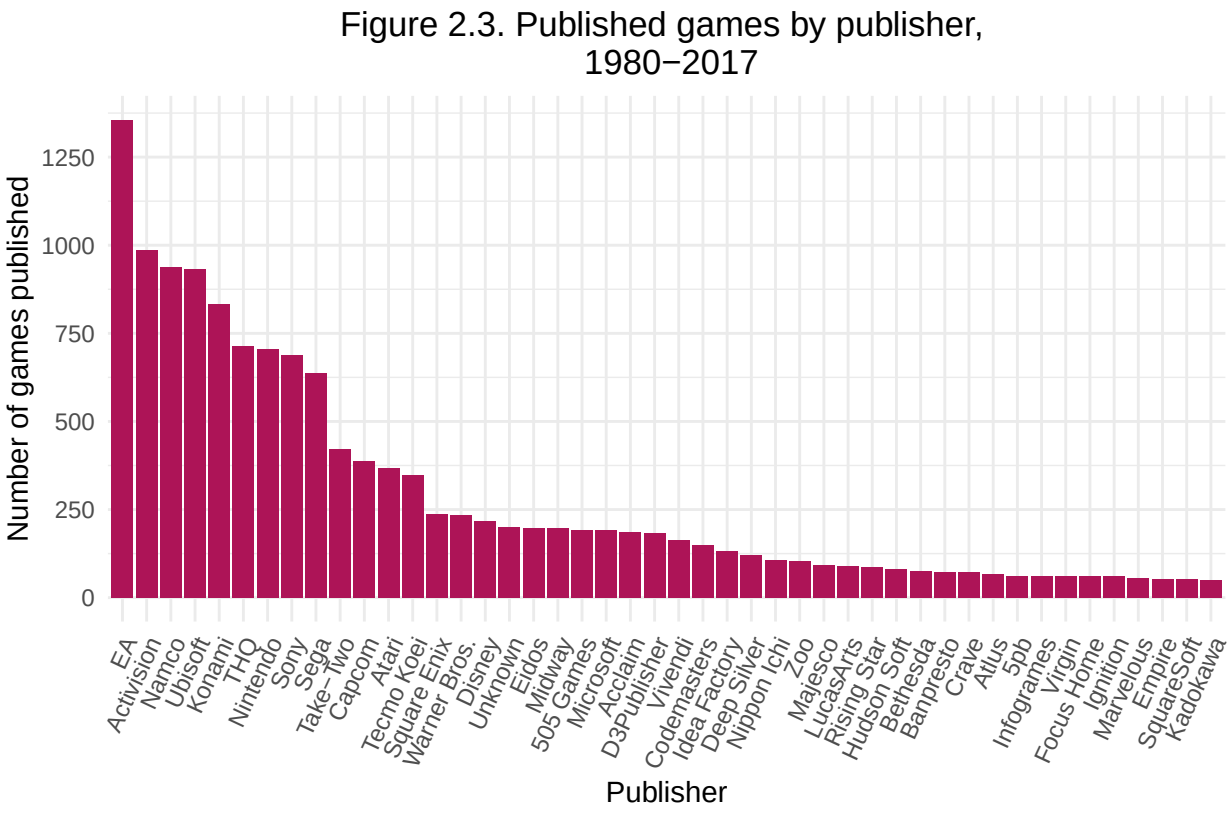
```
length(unique(videogames$Publisher))
```

```
## [1] 581
```

For purposes of plotting, the following graph shows the number of games published by Publisher just for the ones with more that 25 games. Not surprisingly, EA, Activision, Ubisoft and Konami are among the top publishers in this period:

```
videogames %>%
  group_by(Publisher) %>%
  filter(n() >= 50 & !is.na(Publisher)) %>%
  mutate(Publisher = recode(Publisher,
                            'Electronic Arts' = "EA",
                            'Namco Bandai Games' = "Namco",
                            'Konami Digital Entertainment' = "Konami",
                            'Sony Computer Entertainment' = "Sony",
                            'Take-Two Interactive' = "Take-Two",
                            'Warner Bros. Interactive Entertainment' = "Warner Bros.",
                            'Disney Interactive Studios' = "Disney",
                            'Eidos Interactive' = "Eidos",
                            'Midway Games' = "Midway",
                            'Microsoft Game Studios' = "Microsoft",
                            'Acclaim Entertainment' = "Acclaim",
                            'Vivendi Games' = "Vivendi",
                            'Nippon Ichi Software' = "Nippon Ichi",
```

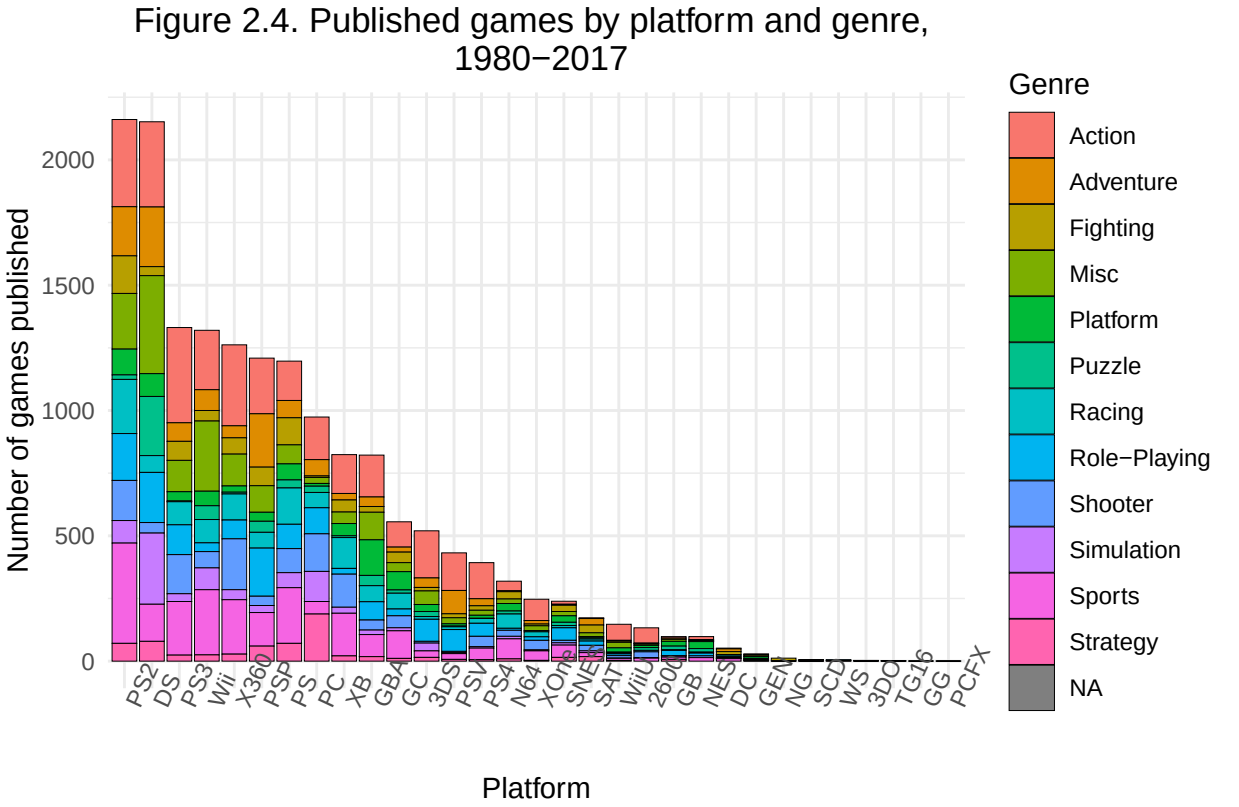
```
                            'Zoo Digital Publishing' = "Zoo",
                            'Majesco Entertainment' = "Majesco",
                            'Rising Star Games' = "Rising Star",
                            'Bethesda Softworks' = "Bethesda",
                            'Crave Entertainment' = "Crave",
                            'Virgin Interactive' = "Virgin",
                            'Focus Home Interactive' = "Focus Home",
                            'Ignition Entertainment' = "Ignition",
                            'Marvelous Interactive' = "Marvelous",
                            'Empire Interactive' = "Empire",
                            'Kadokawa Shoten' = "Kadokawa")) %>%
  ggplot(aes(x = reorder(Publisher, Publisher, function(x) - length(x)))) +
  geom_bar(fill = "#AD1457") +
  scale_y_continuous(breaks = c(seq(0, 1500, 250))) +
  labs(title = "Figure 2.3. Published games by publisher, \n 1980-2017",
       y = "Number of games published",
       x = "Publisher",
       caption = "Source: Own elaboration based on the data 'videogames' from kaggle ") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text.x = element_text(angle = 65, hjust = 1), plot.caption = element_text(hjust= 0))
```



Giving those insights, one can plot the games published both by platform and by genre:

```
videogames %>%
  ggplot(aes(x = reorder(Platform, Platform, function(x) - length(x)), fill = Genre)) +
```

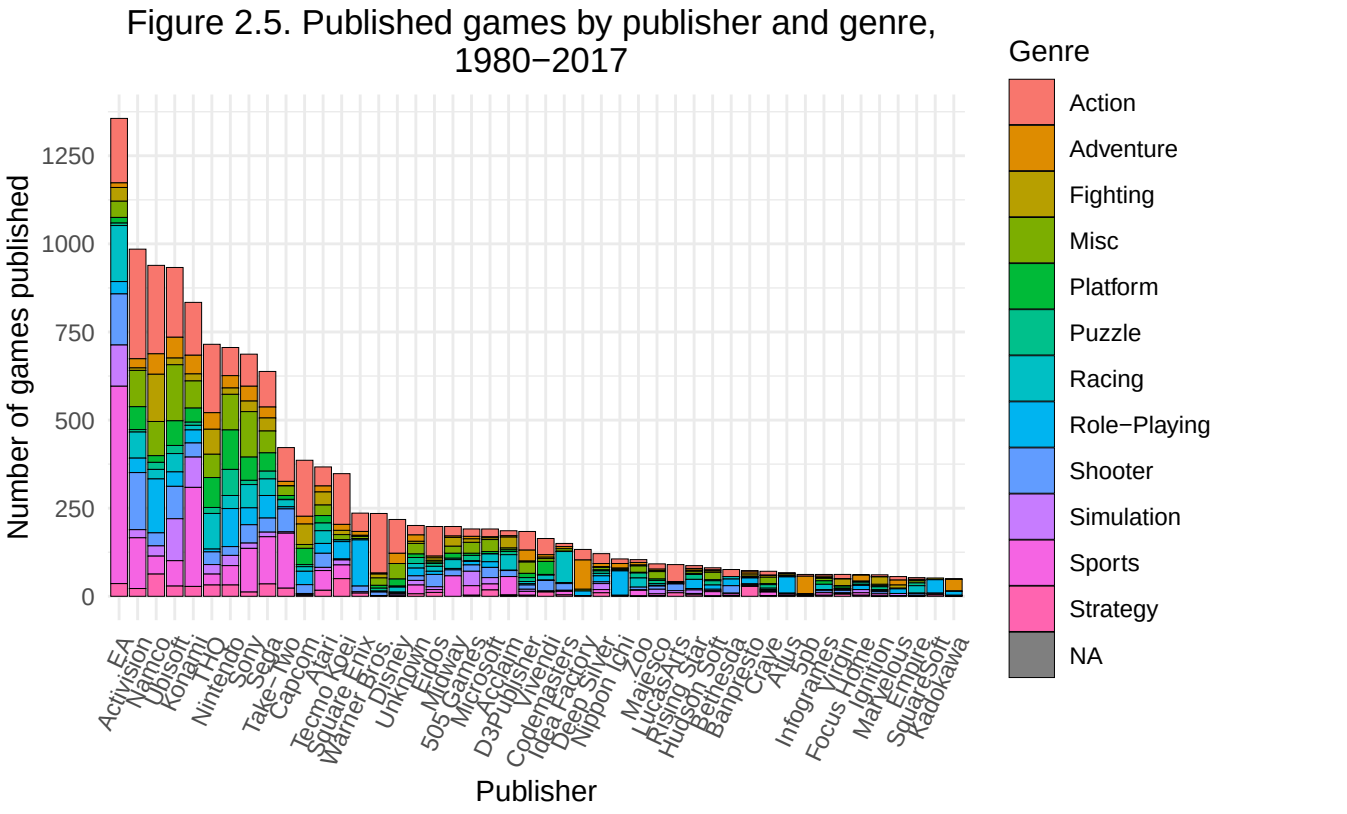
```
  geom_bar(color = "black", size = 0.1) +
  labs(title = "Figure 2.4. Published games by platform and genre, \n 1980-2017",
       y = "Number of games published",
       x = "Platform",
       caption = "Source: Own elaboration based on the data 'videogames' from kaggle ") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text.x = element_text(angle = 65, hjust = 0), plot.caption = element_text(hjust= 0))
```



In general, there is no clear pattern of a platform with a tendency to publish more or less specific genres. Does this hold for publishers and genres?

```
videogames %>%
  group_by(Publisher) %>%
  filter(n() >= 50 & !is.na(Publisher)) %>%
  mutate(Publisher = recode(Publisher,
                            'Electronic Arts' = "EA",
                            'Namco Bandai Games' = "Namco",
                            'Konami Digital Entertainment' = "Konami",
                            'Sony Computer Entertainment' = "Sony",
                            'Take-Two Interactive' = "Take-Two",
                            'Warner Bros. Interactive Entertainment' = "Warner Bros.",
                            'Disney Interactive Studios' = "Disney",
                            'Eidos Interactive' = "Eidos",
                            'Midway Games' = "Midway",
                            'Microsoft Game Studios' = "Microsoft",
```

```
                          'Acclaim Entertainment' = "Acclaim",
                          'Vivendi Games' = "Vivendi",
                          'Nippon Ichi Software' = "Nippon Ichi",
                          'Zoo Digital Publishing' = "Zoo",
                          'Majesco Entertainment' = "Majesco",
                          'Rising Star Games' = "Rising Star",
                          'Bethesda Softworks' = "Bethesda",
                          'Crave Entertainment' = "Crave",
                          'Virgin Interactive' = "Virgin",
                          'Focus Home Interactive' = "Focus Home",
                          'Ignition Entertainment' = "Ignition",
                          'Marvelous Interactive' = "Marvelous",
                          'Empire Interactive' = "Empire",
                          'Kadokawa Shoten' = "Kadokawa")) %>%
ggplot(aes(x = reorder(Publisher, Publisher, function(x) - length(x)), fill = Genre)) +
geom_bar(color = "black", size=0.1) +
scale_y_continuous(breaks = c(seq(0, 1500, 250))) +
labs(title = "Figure 2.5. Published games by publisher and genre, \n 1980-2017",
     y = "Number of games published",
     x = "Publisher",
     caption = "Source: Own elaboration based on the data 'videogames' from kaggle ") +
theme_minimal() +
theme(plot.title = element_text(hjust = 0.5)) +
theme(axis.text.x = element_text(angle = 65, hjust = 1), plot.caption = element_text(hjust= 0))
```



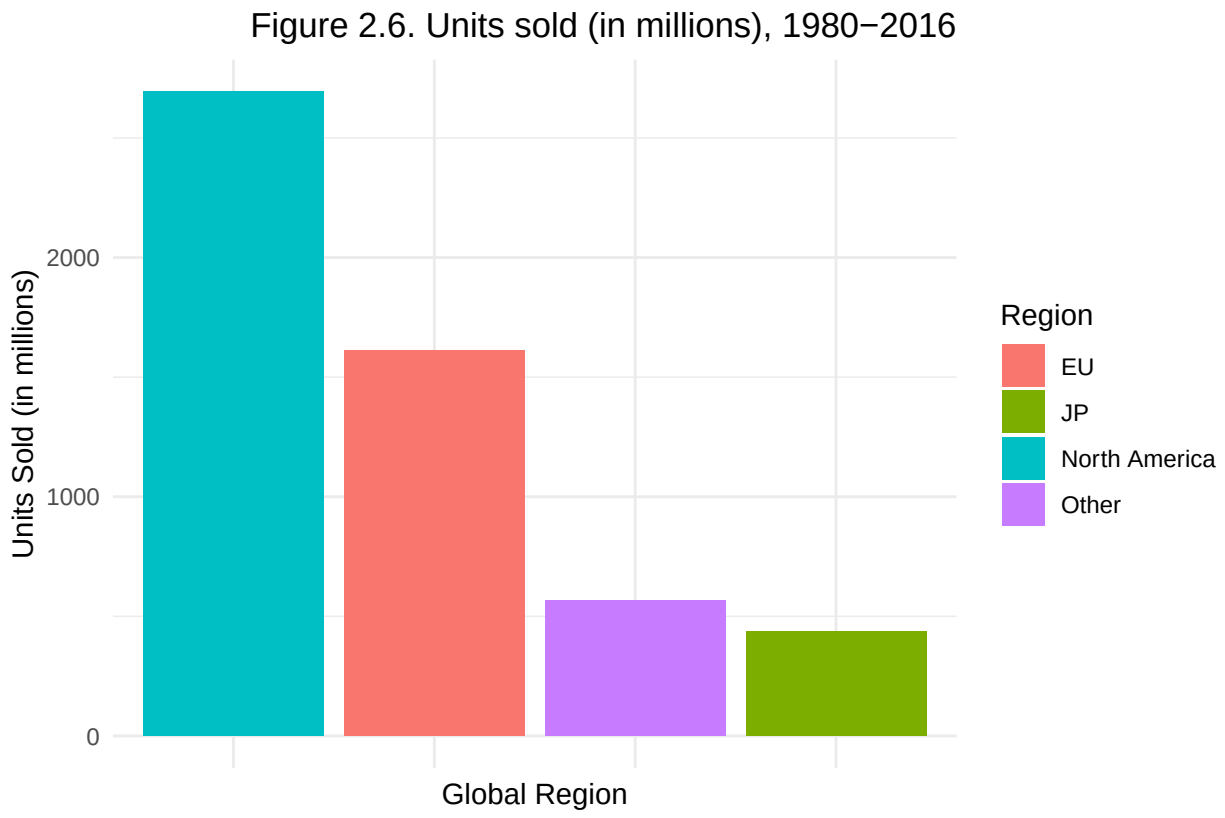
Here, there are some publishers that tend to publish certain genres. For example, EA and Konami have a

lot of sports games; Activision and Namco have a good amount of action games; the majority of Warner Bros. games are action ones; Idea Factory almost only publishes adventure games.

The next set of variables worth noticing is the sales; as it is clear, North America is the Region with more units sold during this period:

```
# for the next graph and the next analysis, the data set was modified as follows

videogames1 <- videogames %>%
#drop NAs
  drop_na() %>%
#drop columns that won't be used
  select(-Critic_Count, -User_Count) %>%
#rename "Years"
  rename(Year = Year_of_Release) %>%
# coerce variables as correct objects
  mutate(Platform = as.factor(Platform),
         Genre = as.factor(Genre),
         Publisher = as.factor(Publisher),
         Developer = as.factor(Developer),
         Rating = as.factor(Rating),
         User_Score = as.numeric(User_Score) * 10)

#filter blank ratings and developers
videogames1 <- videogames1 %>%
  filter(Rating != "" & Developer != "")

videogames1 <- videogames1 %>%
# correctly rename Sales
  pivot_longer(cols = ends_with("Sales"),
                                names_to = "Region", values_to = "Units_Sold")

# rename NA region
videogames1$Region = gsub("NA_Sales", "North America", videogames1$Region)

# consolidate Sales
videogames1$Region = gsub("_Sales", "", videogames1$Region)
```

```
videogames1 %>%
  filter(Region != "Global") %>%
  ggplot(aes(reorder(Region, -Units_Sold), y = Units_Sold)) +
  geom_bar(aes(fill = Region), stat = "identity") +
  labs(x = "Global Region",
       y = "Units Sold (in millions)",
       title = "Figure 2.6. Units sold (in millions), 1980-2016",
       caption = "Source: Own elaboration based on the data 'videogames' from kaggle") +
  theme_minimal() +
  theme(axis.text.x = element_blank()) +
  theme(plot.title = element_text(hjust = 1), plot.caption = element_text(hjust= 0))
```



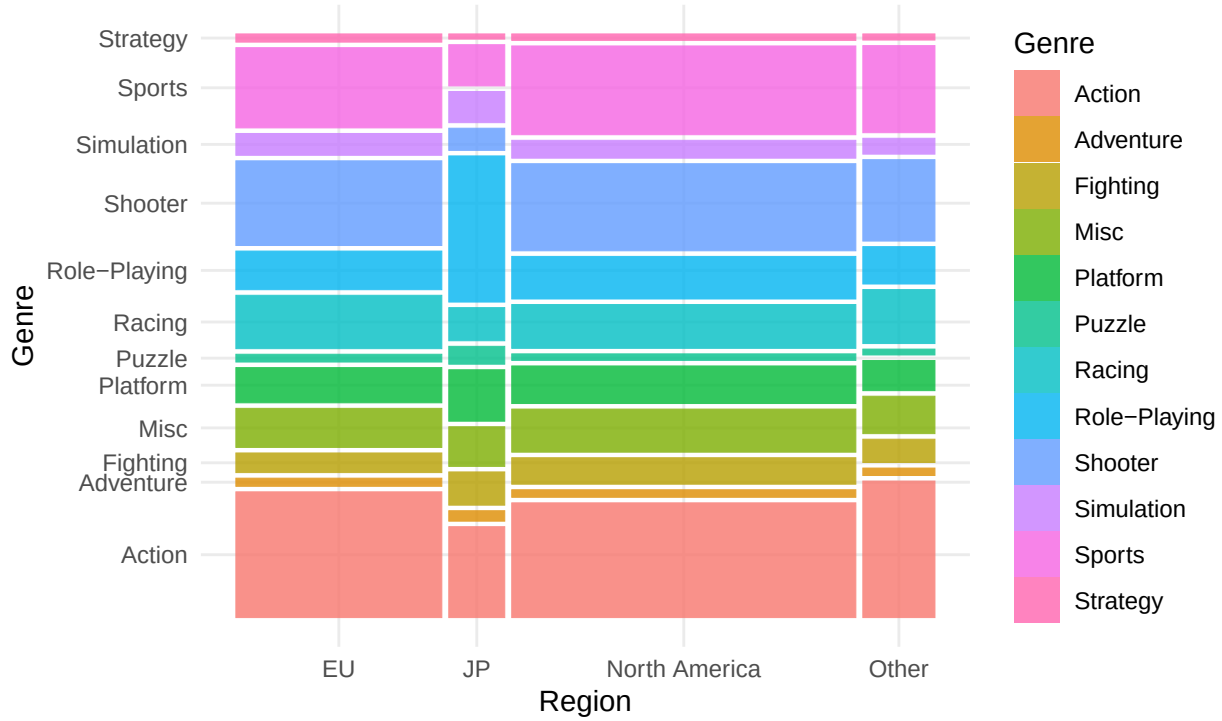
Source: Own elaboration based on the data 'videogames' from kaggle

But how does this composition changed over time? The next graph shows the composition of games sales by region. It is clear that in the last 20 years, North America is the region with the largest volume of sales, followed by the European Union. Despite being a small proportion of total sales, other regions have increased their participation as seen below:

Another relevant question is to ask how the sales are distributed by genre within each region? The next couple of graphs show how Europe and North America disproportionately prefer Sports and Shooter games over other regions. On the other hand, Japanese user prefer Role-Playing games, while the rest of the world likes action more.

```
videogames1 %>%
  filter(Region != "Global") %>%
  ggplot() +
  geom_mosaic(aes(product(Genre, Region),
                  fill = Genre, weight = Units_Sold)) +
  labs(x = "Region", y = "Genre",
       title = "Figure 2.7. Composition of regional sales by genre, \n 1980-2017",
       caption = "Source: Own elaboration based on the data 'videogames' from kaggle ") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5), plot.caption = element_text(hjust= 0))
```

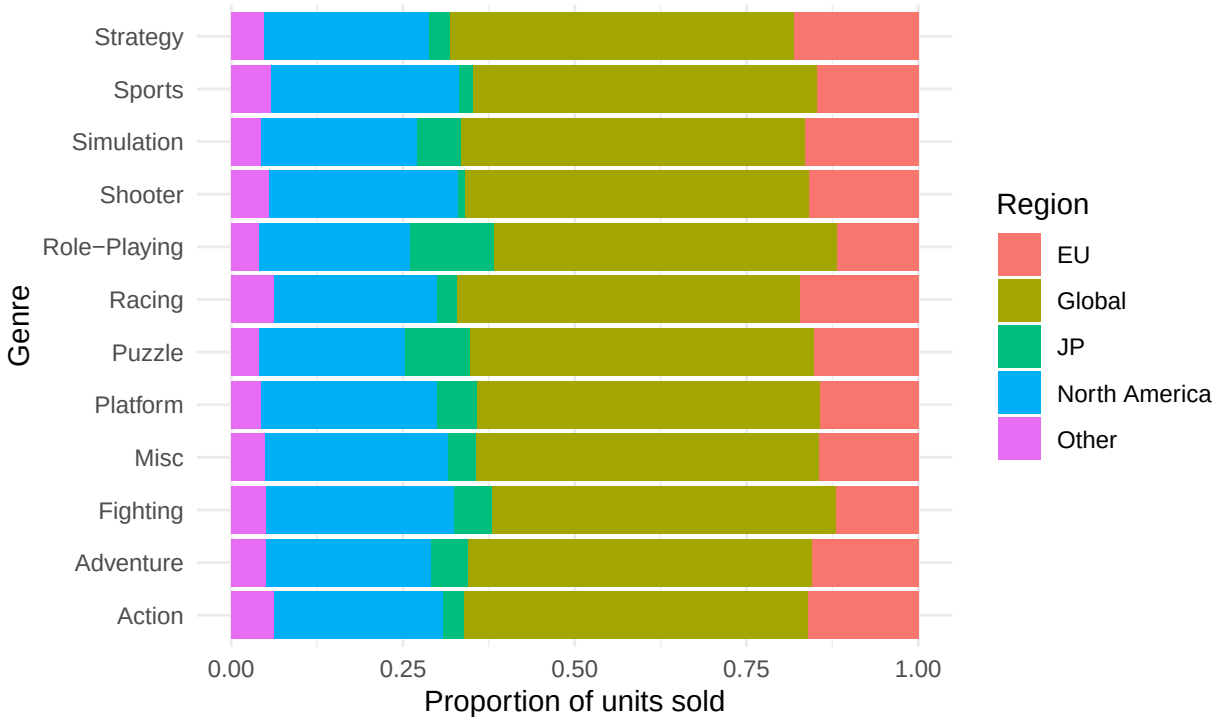
Figure 2.7. Composition of regional sales by genre, 1980−2017



Source: Own elaboration based on the data 'videogames' from kaggle

```
videogames1 %>%
  filter(Region != "Global_Sales") %>%
  ggplot(aes(x = Genre, y = Units_Sold)) +
  geom_bar(aes(fill = Region), stat = "identity", position = "fill") +
  ylab("Proportion of units sold") +
  coord_flip() +
  labs(title = "Figure 2.8. Proportion of units sold by genre and region, \n 1980, 2017",
       caption = "Source: Own elaboration based on the data 'videogames' from kaggle ") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5), plot.caption = element_text(hjust= 0))
```

Figure 2.8. Proportion of units sold by genre and region, 1980, 2017



Source: Own elaboration based on the data 'videogames' from kaggle

3 Linear Regression Models

3.1 User and Critic Scores

In this section, the relationship between user and critics scores and sales is analyzed. Firstly, lets observe how both scores are related. One could think that they will have a pretty high correlation. If we fit a model to explain the User Score with the Critic Score, the following results are obtained:

Table 3.1. Linear model, User Score vs. Critic Score

	User Score
Critic Score	0.603*** (0.005)
Constant	29.515*** (0.328)
Observations	34,125
R^2	0.337
Adjusted R^2	0.337
Residual Std. Error	11.726 (df = 34123)
F Statistic	17,326.680*** (df = 1; 34123)
Significance Value (P-value)	*p<0.1; **p<0.05; ***p<0.01

These results imply the following estimated model:

$$\hat{UserScore} = 29.515 + 0.603 CriticScore + \varepsilon$$

This means that the Critic Score is highly significant to explain the behavior of the User Score, with a significance level of 99%. Additionally, the standard error is very small. Lets now compute the inverse relationship: explain Critic Score by USer Score:

Table 3.2. Linear model, Critic Score vs. User Score

	Critic Score
User Score	0.559*** (0.004)
Constant	30.110*** (0.311)
Observations	34,125
R^2	0.337
Adjusted R^2	0.337
Residual Std. Error	11.294 (df = 34123)
F Statistic	17,326.680*** (df = 1; 34123)
Significance Value (P-value)	*p<0.1; **p<0.05; ***p<0.01

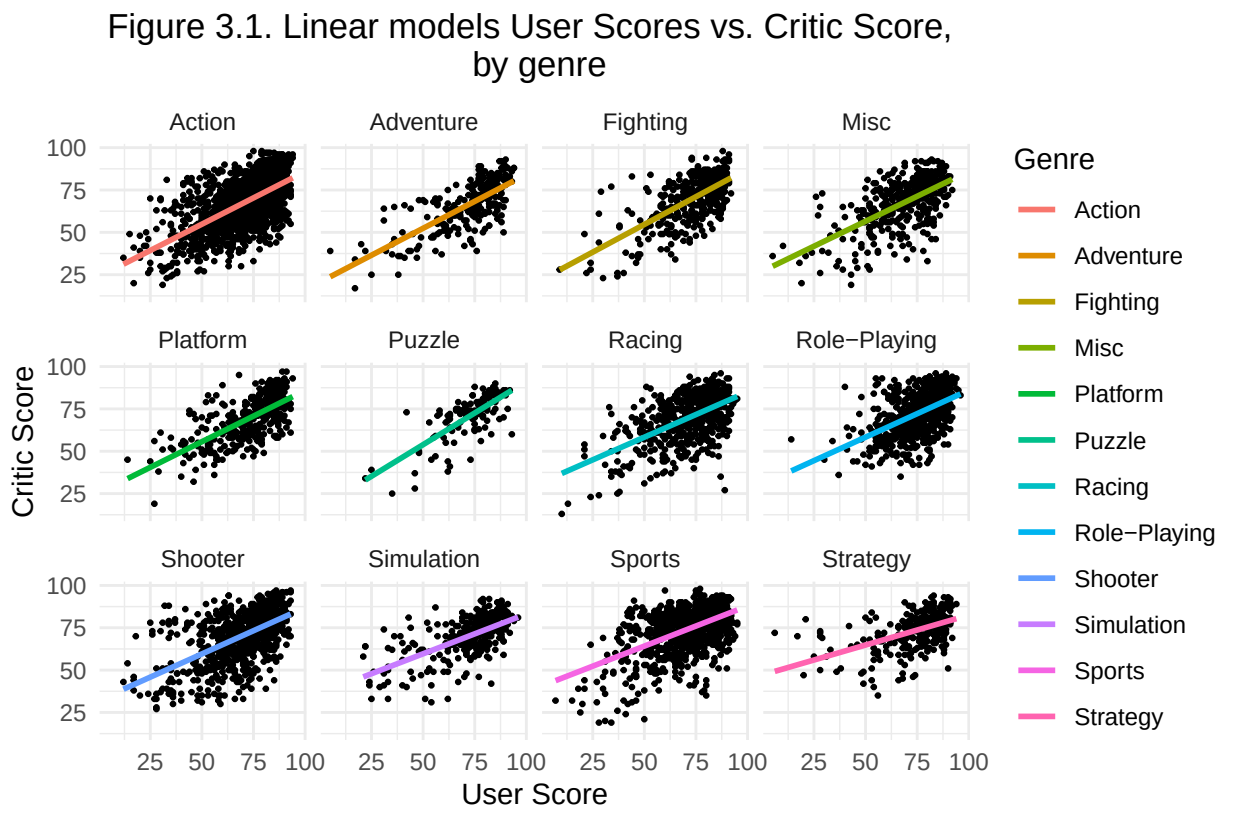
Which implies:

$$\hat{CriticScore} = 30.110 + 0.559 UserScore + \varepsilon$$

Not surprisingly, the results are highly similar to the previous ones. Thus, when addressing the relationship between both scores, it is sensible to assure a strong association and correlation between the two variables (note also that the R^2 is relatively substantial and should not be disregarded.

Nonetheless, the aforementioned models are somehow naive. It would be better to analyze the relationship adding another element of variable. Because our exploratory data analysis suggested interesting patterns within genres, lets facet the linear model with this variable:

```
videogames1 %>%
  filter(Region == "Global") %>%
  ggplot(aes(x = User_Score, y = Critic_Score)) +
    geom_point(size = 0.5) +
    geom_smooth(aes(color = Genre), method = lm, se = FALSE) +
    facet_wrap(~ Genre) +
    xlab("User Score") +
    ylab("Critic Score") +
    theme_minimal() +
    labs(title = "Figure 3.1. Linear models User Scores vs. Critic Score, \n by genre",
         caption = "Source: Own elaboration based on the data 'videogames' from kaggle ") +
    theme(plot.title = element_text(hjust = 0.5), plot.caption = element_text(hjust= 0))
```



The results above strengthen the assertion of a strong positive relationship between both scores. Despite this, it is clear that the relationship varies across genres. It can be seen that, for example, Simulation and Strategy games have a weaker relationship between score than adventure and puzzle ones.

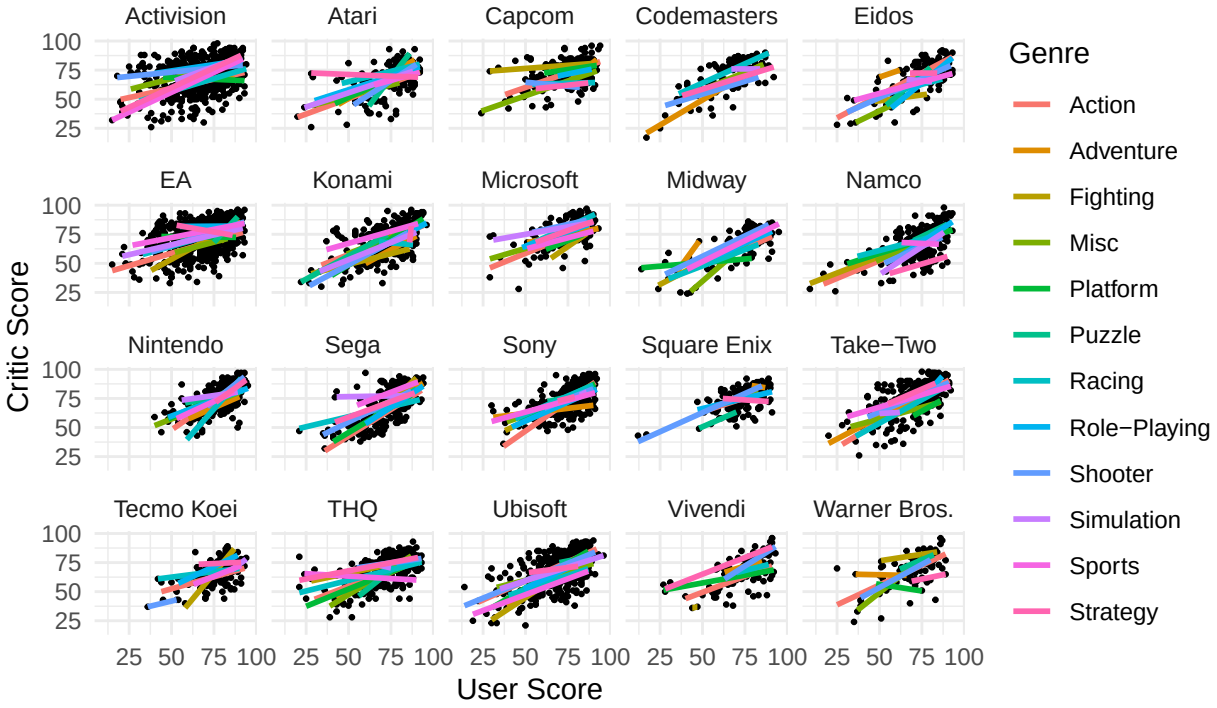
Concluding this analysis, one can say that higher user scores are correlated with higher critic scores, and viceversa. Although this correlation varies across genres, it is consistent and always positive.

3.2 User/Critic Scores by Genre and Publisher

Now lets fit the same model but faceting by publishers with more than 500 games pudblished:

```
videogames1 %>%
  group_by(Publisher) %>%
  mutate(Publisher = recode(Publisher,
                            'Electronic Arts' = "EA",
                            'Namco Bandai Games' = "Namco",
                            'Take-Two Interactive' = "Take-Two",
                            'Konami Digital Entertainment' = "Konami",
                            'Sony Computer Entertainment' = "Sony",
                            'Warner Bros. Interactive Entertainment' = "Warner Bros.",
                            'Eidos Interactive' = "Eidos",
                            'Midway Games' = "Midway",
                            'Microsoft Game Studios' = "Microsoft",
                            'Vivendi Games' = "Vivendi")) %>%
  filter(n() >= 500 & Region == "Global" ) %>%
  ggplot(aes(x = User_Score, y = Critic_Score)) +
    geom_point(size = 0.5) +
    geom_smooth(aes(color = Genre), method = lm, se = FALSE) +
    facet_wrap(~ Publisher) +
    xlab("User Score") +
    ylab("Critic Score") +
    theme_minimal() +
    labs(title = "Figure 3.2. Linear models User Scores vs. Critic Score, \n by genre and publisher",
         caption = "Source: Own elaboration based on the data 'videogames' from kaggle ") +
    theme(plot.title = element_text(hjust = 0.5), plot.caption = element_text(hjust= 0))
```

Figure 3.2. Linear models User Scores vs. Critic Score, by genre and publisher



Source: Own elaboration based on the data 'videogames' from kaggle

Now, the relationships are very broad, so consider the following cases:

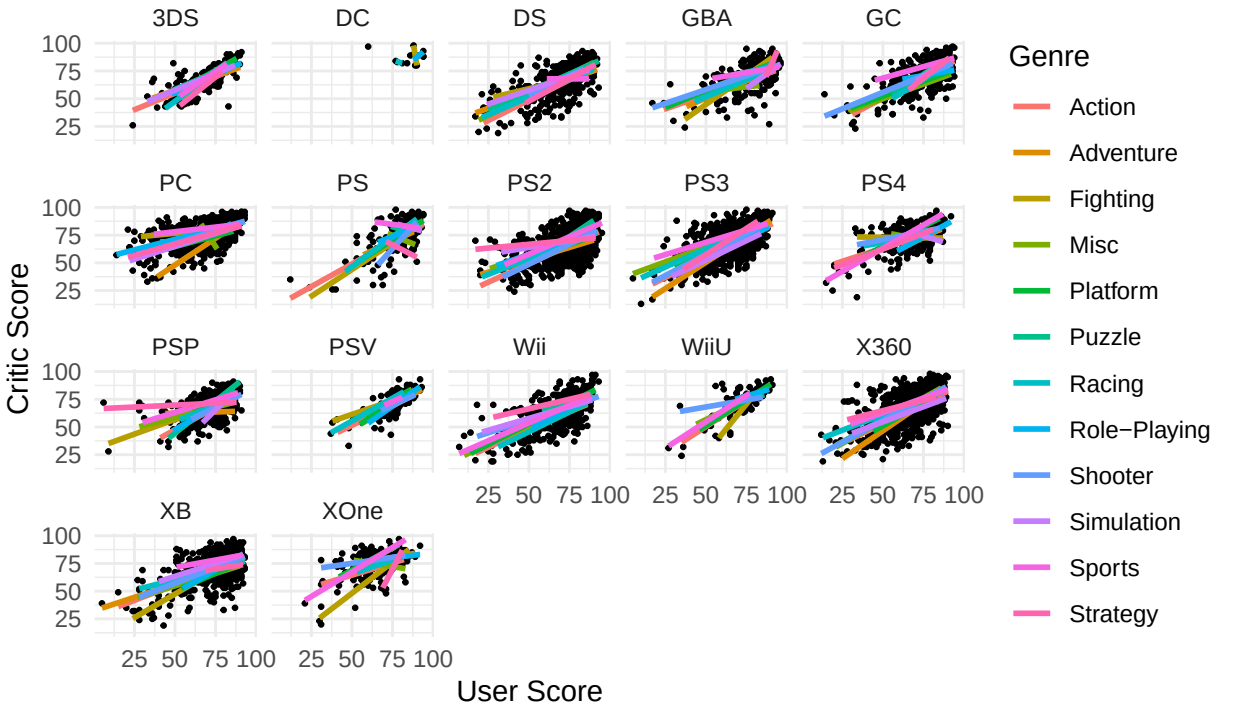
1. Publishers with all/almost all correlations clearly positives. Publishers like Codemasters, Konami, Microsoft, Sony, Take-Two, Ubisoft and Vivendi show a strong relationship between user and score critics across all genres.
2. Publishers with clearly some genres with much weaker correlations. Activision, Atari, Eidos, EA, Midway, Namco, Nintendo and Sega clearly show at least one weak relationship between both scores, suggesting that these genres may have lower User Scores that do not explain the Critic Score behavior. Nonetheles, the aggregate of all genres still shows a strong positive correlation.
3. Publishers with weak overall correlations. Capcom, Square Enix, Tecmo Koei, THQ and Warner Bros. are publishers whose relationships between scores are pretty weak. Despite showing some positive correlations in some genres, the aggregated behavior is messier.

3.3 User/Critic Scores by Genre and Platform

The following graphs replicate the procedure above but faceting by platform:

```
videogames1 %>%
  filter(Region == "Global") %>%
  ggplot(aes(x = User_Score, y = Critic_Score)) +
    geom_point(size = 0.5) +
    geom_smooth(aes(color = Genre), method = lm, se = FALSE) +
    facet_wrap(~ Platform) +
    xlab("User Score") +
    ylab("Critic Score") +
    theme_minimal() +
    labs(title = "Figure 3.3. Linear models User Scores vs. Critic Score, \n by genre and platform",
         caption = "Source: Own elaboration based on the data 'videogames' from kaggle ") +
    theme(plot.title = element_text(hjust = 0.5), plot.caption = element_text(hjust= 0))
```

Figure 3.3. Linear models User Scores vs. Critic Score, by genre and platform



Source: Own elaboration based on the data 'videogames' from kaggle

Again, some different patterns are clear: firstly, platform DC has too few games so a proper analysis can be made; second, almost all platforms show explicit and strong correlations between scores; finally, both WiiU and XBox One have somehow messier correlations, where some genres show very weak correlations. However, in general, all platforms (except DC) clearly positive aggregated relationships.

3.4 User/Critic Scores and Units Sold

Perhaps the most relevant insight from the data is to study the relationship between scores and sales. One could argue that higher User/Critic scores will lead to higher sales for any game. Firstly, lets fit two aggregated models that accounts for sales and User/Critic scores:

Table 3.3. Linear model, Units Sold vs. User Score

	Units Sold
User Score	0.005*** (0.0004)
Constant	−0.035 (0.030)
Observations	34,125
R^2	0.004
Adjusted R^2	0.004
Residual Std. Error	1.071 (df = 34123)
F Statistic	143.184*** (df = 1; 34123)
Significance Value (P-value)	*p<0.1; **p<0.05; ***p<0.01

Which implies a model:

$$\hat{UnitsSold} = -0.035 + 0.005 UserScore + \varepsilon$$

The table above could appear surprising, since it suggests that user scores do not explain almost anything of the units sold: despite being highly significant (at a 99% level) the coefficient and the R^2 are way too low to assure that our predictor is in fact significant for our dependent variable.

Nevertheless, it is important to consider the following: user scores are sometimes made long time after the games are released, so the units sold do not clearly occur after user scores. Lets examine the opposite relationship: explain the user score with the units sold:

Table 3.4. Linear model, User Score vs. Units Sold

	User Score
Units Sold	0.867*** (0.072)
Constant	71.587*** (0.081)
Observations	34,125
R^2	0.004
Adjusted R^2	0.004
Residual Std. Error	14.369 (df = 34123)
F Statistic	143.184*** (df = 1; 34123)
Significance Value (P-value)	*p<0.1; **p<0.05; ***p<0.01

This table implies the next model:

$$\hat{UserScore} = 71.587 + 0.867 UnitsSold + \varepsilon$$

Now, the correlation is pretty high, which suggests that one could explain the units sold by the user scores. However, the relationship seems counter intuitive, and the R^2 proves that: it is very small, which means that, although the independent variable is correlated with the dependent variable, the model does not explain almost nothing of the variability of the latter one.

Lets now turn the attention to two models that relate critic scores and units sold:

Table 3.5. Linear model, Critic Score vs. Units Sold

	Critic Score
Units Sold	2.245*** (0.069)
Constant	69.574*** (0.077)
Observations	34,125
R^2	0.030
Adjusted R^2	0.030
Residual Std. Error	13.657 (df = 34123)
F Statistic	1,062.868*** (df = 1; 34123)
Significance Value (P-value)	*p<0.1; **p<0.05; ***p<0.01

Unlike the first model (User Score ~ Units Sold), the one above shows a clearly strong and significant correlation between critic scores and units sold. Thus, with a significance level of 99%, it is reasonable to assure that a higher amount of units sold can increase the critic score by an average value of 2.245. The model would look like this:

$$\hat{CriticScore} = 69.574 + 2.245 UnitsSold + \varepsilon$$

Lets invert the relationship:

Table 3.6. Linear model, Units Sold vs. Critic Score

	Units Sold
Critic Score	0.013*** (0.0004)
Constant	−0.634*** (0.030)
Observations	34,125
R^2	0.030
Adjusted R^2	0.030
Residual Std. Error	1.057 (df = 34123)
F Statistic	1,062.868*** (df = 1; 34123)
Significance Value (P-value)	*p<0.1; **p<0.05; ***p<0.01

Again, a significant estimated coefficient is shown; however, once more it is very small, while the R^2 is also expendable:

$$\hat{UnitsSold} = -0.634 + 0.013 CriticScore + \varepsilon$$

For the next model, lets fit both scores to explain the units sold, as follows:

Table 3.7. Linear model, Units Sold vs. Critic Score + User Score

	Units Sold
Critic Score	0.016*** (0.001)
User Score	−0.004*** (0.0005)
Constant	−0.514*** (0.033)
Observations	34,125
R^2	0.032
Adjusted R^2	0.032
Residual Std. Error	1.056 (df = 34122)
F Statistic	567.371*** (df = 2; 34122)
Significance Value (P-value)	*p<0.1; **p<0.05; ***p<0.01

As a general model, this shows that neither score has a practically large impact on units sales (although both are highly significant); additionally, the variance of the dependent variable is almost not explained by the predictive ones. This model would look as follows:

$$\hat{UnitsSold} = -0.514 + 0.016CriticScore + -0.004UserScore + \varepsilon$$

Now, a faceting by genre will be made in order to show how both scores relate to the units sold in each case:

```
videogames1 %>%
  filter(Region == "Global") %>%
  pivot_longer(cols = c(User_Score, Critic_Score), names_to = "Rating_Type",
                        values_to = "Score") %>%
  ggplot(aes(x = Score, y = Units_Sold)) +
    geom_smooth(aes(color = Rating_Type), method = "lm", se = FALSE) +
    xlab("User and Critic Scores") +
    ylab("Units Sold") +
    labs(title = "Figure 3.4. Linear model, User/Critic Scores vs. Units Sold, \n by genre",
            caption = "Source: Own elaboration based on the data 'videogames' from kaggle ") +
    facet_wrap(~ Genre) +
    theme_minimal() +
    theme(plot.title = element_text(hjust = 0.5), plot.caption = element_text(hjust= 0))
```

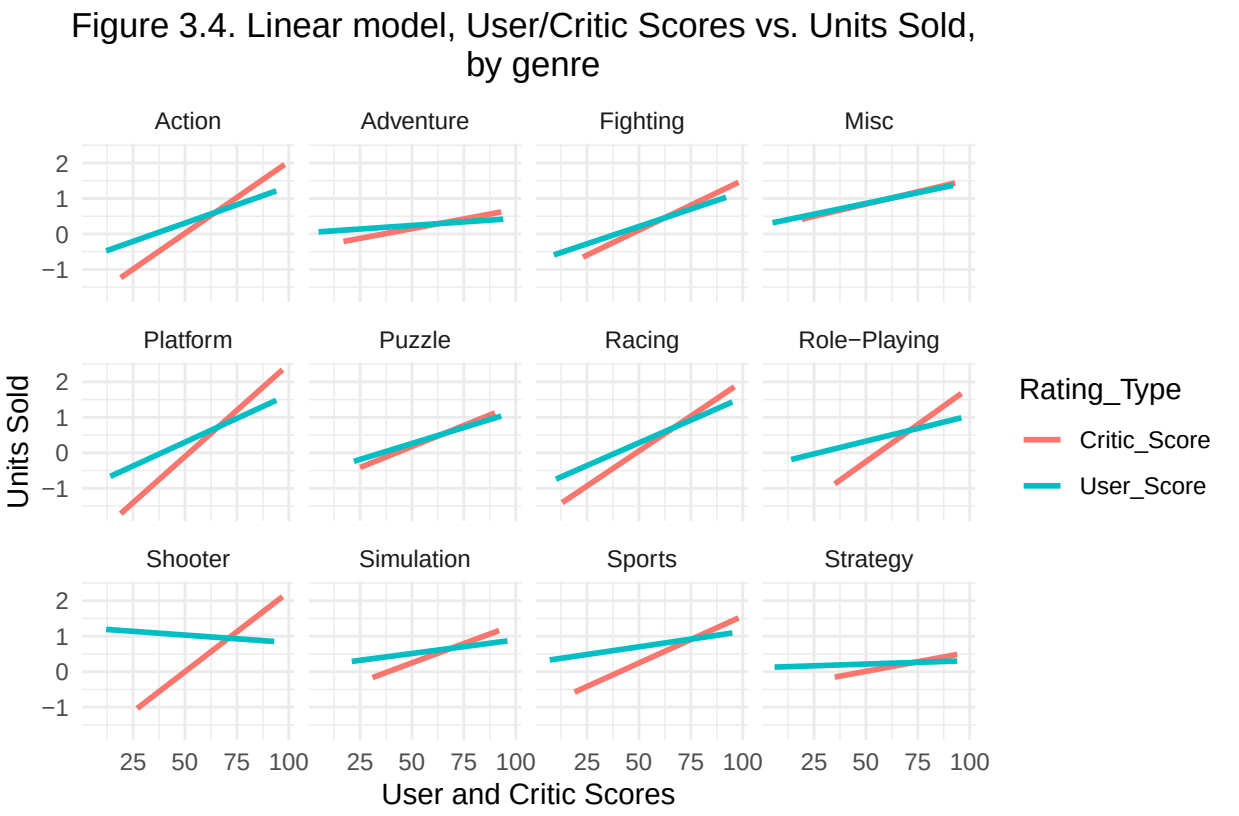


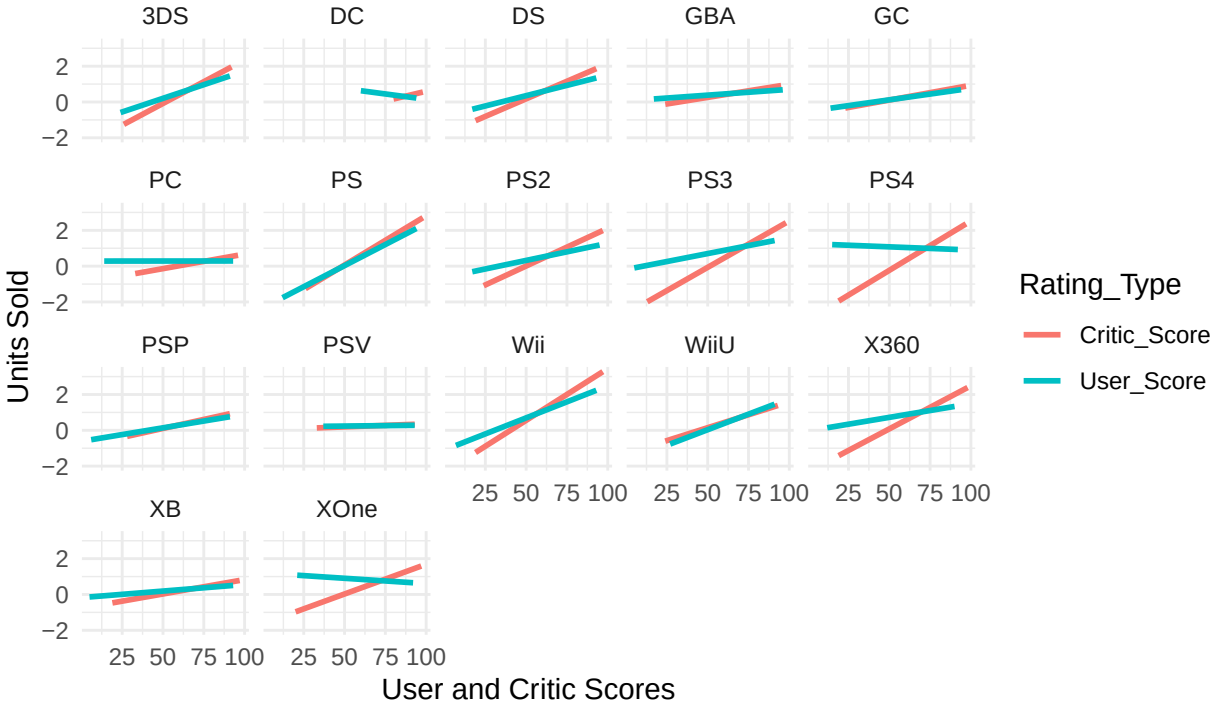
Figure 2.4. shows very meaningful insight about the data: first, in almost all genres (except Misc), Critic Scores display a steeper slope, suggesting a stronger relationship with the units sold (which is consistent with previous results). As a consequence, one can say that games consumers are more influenced by critic scores than by other user's.

Second, some genres depict very weak correlations for both types of scores: adventure, misc and strategy are some of them. Intuitively, games buyers for this games do no take too much into account the game ratings, regardless of whether they come from critics or users. In fact, particular cases shooter games show that user ratings are not correlated at all with units sold, while critic scores display a high slope.

How does this faceting behave under Platform?

```
videogames1 %>%
  filter(Region == "Global") %>%
  pivot_longer(cols = c(User_Score, Critic_Score), names_to = "Rating_Type",
                          values_to = "Score") %>%
  ggplot(aes(x = Score, y = Units_Sold)) +
    geom_smooth(aes(color = Rating_Type), method = "lm", se = FALSE) +
    xlab("User and Critic Scores") +
    ylab("Units Sold") +
    labs(title = "Figure 3.5. Linear model, User/Critic Scores vs. Units Sold \n by Platform",
            caption = "Source: Own elaboration based on the data 'videogames' from kaggle ") +
    facet_wrap(~ Platform) +
    theme_minimal() +
    theme(plot.title = element_text(hjust = 0.5), plot.caption = element_text(hjust= 0))
```

Figure 3.5. Linear model, User/Critic Scores vs. Units Sold by Platform



Source: Own elaboration based on the data 'videogames' from kaggle

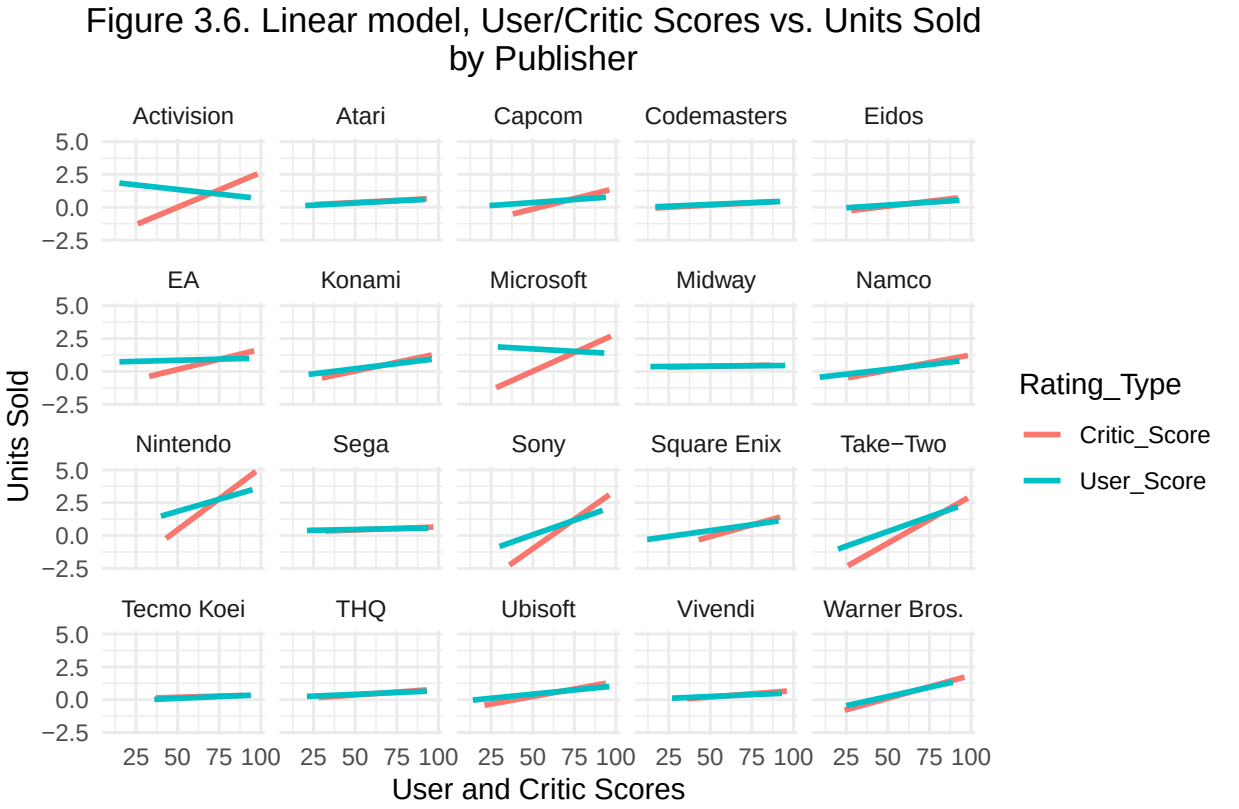
One main insight arises: more platforms have weaker relationships between user or critic scores and units sold. When faceting by genre, the majority of the correlations were clearly and strongly positive; in this case, several platforms show more horizontal slopes (there even are two negatives). Another pattern repeats, where critic scores display a stronger correlation with units sold than the user ones (in all cases, the critic score slope is steeper).

This also assures the previous models where critic scores were more explanatory about the units sold.

Finally, consider a faceting by publisher:

```
videogames1 %>%
  group_by(Publisher) %>%
  mutate(Publisher = recode(Publisher,
                            'Electronic Arts' = "EA",
                            'Namco Bandai Games' = "Namco",
                            'Take-Two Interactive' = "Take-Two",
                            'Konami Digital Entertainment' = "Konami",
                            'Sony Computer Entertainment' = "Sony",
                            'Warner Bros. Interactive Entertainment' = "Warner Bros.",
                            'Eidos Interactive' = "Eidos",
                            'Midway Games' = "Midway",
                            'Microsoft Game Studios' = "Microsoft",
                            'Vivendi Games' = "Vivendi")) %>%
  filter(n() >= 500 & Region == "Global") %>%
  pivot_longer(cols = c(User_Score, Critic_Score), names_to = "Rating_Type",
                          values_to = "Score") %>%
  ggplot(aes(x = Score, y = Units_Sold)) +
    geom_smooth(aes(color = Rating_Type), method = "lm", se = FALSE) +
```

```
    xlab("User and Critic Scores") +
    ylab("Units Sold") +
    labs(title = "Figure 3.6. Linear model, User/Critic Scores vs. Units Sold \n by Publisher",
           caption = "Source: Own elaboration based on the data 'videogames' from kaggle ") +
    facet_wrap(~ Publisher) +
    theme_minimal() +
    theme(plot.title = element_text(hjust = 0.5), plot.caption = element_text(hjust= 0))
```



Once more, user scores appear to be less relevant for the consumers' decisions; additionally, critic scores are more explanatory of the units sold than the user scores. In general, both scores show a weak correlation with units sold across a lot of publishers; just some of them (Microsoft, Nintendo, Sony, and Take-Two) display a explicit positive relationship between the variables.

4 Conclusions

Throughout this document, several data descriptions were made, where insight was gained about the platforms, genres and publisher with more games published. Additionally, some insights about the game genre composition for the top publishers and platforms showed tendencies for some of them to publish more or less games of specific genres.

Furthermore, and not surprisingly, it was revealed that North America and Europe were the regions with the largest game markets, while Japan and the rest of the world represented smaller amount of units sold. It was also shown that certain regions have a preference for some game genres over others.

When fitting linear models to study several relationships between units sold, scores, genres, publishers and platforms, there are two mean conclusions than can be assured:

1. Scores or ratings does partially matter to predict units sold. Specifically, critic scores showed a consistent stronger positive effect to explain it, while user scores displayed a much more inconsistent behavior.

Considering the first hypothesis raised in the Introduction, the models resulted in disparate behaviors between user and critic scores, where the last one consistently appeared more predictive and significant.

One possible explanation for this is that maybe that consumers value more that ratings of the called “experts” rather than the ones of fellow users.

Speaking of genre, consumers clearly value opinions more in some cases (like Action, Fighting, Platform, Racing, Role-Playing, Shooter and Sports); here something is key: this does not depend of the total amount of units sold by genre within the studied period.

On the contrary, the commercial success for other genres does not depend on neither user nor critic scores. This could mean that consumers of this genres care about a lot of other elements before the ratings.

In terms of platforms, something expected showed up: consumers of the biggest console brands (PlayStation and XBox) did care about critic scores (although not so much about user ones). The one case which represents a major surprise is PC, where user scores explain almost nothing of units sold (the slope is practically horizontal) and the critic scores are lightly positive, but very weak. Since the PC is a huge games market, one could suggest to the game industry to pay attention to another elements other than ratings to sell their game copies.

In the case of publishers, it was showed how some of them strongly rely on critic scores to sell copies (Activision, EA, Microsoft, Nintendo, Sony and Take-Two). Could this be an indicator that, for this publishers, consumers wait for good reviews on a game to buy it, given bad experiences in the past? This hypothesis would imply that publishers with weak correlations with units sold are more consistent in their game quality. However, that’s the direction for future investigations.

2. Therefore, in an aggregate level, one cannot say that scores clearly explain the units sold. The general models fitted, though with significant coefficients, explained a very little amount of the dependent variable variance and the effect of that coefficient was effectively expendable.

However, the rating did explain some cases. It depended. On what? As it was shown, on several variables like the score source, genre, platform and publisher. Surely, there are more than could be used to deepen the analysis, but the scope and objectives of this document are fulfilled. Therefore, the second hypothesis is partially correct, since some variables are in fact positively correlated with units sold, while other are not.

5 References

HarvardX data science professional certificate. (n.d.). In *edX*. https://www.edx.org/es/professional-certificate/harvardx-data-science?index=spanish_product&queryID=5be4100055746931457d0566e5c3bf26&position=2

Shiny project. (2021). In *GitHub*. Casey Hoffman. https://github.com/caseyahoffman/shinyproject/tree/main/shinyapp

Video game industry - statistics & facts. (2021). In *Statista*. J. Clement. https://www.statista.com/topics/868/video-games/