Information Visualization

Video Game

*By* Meiru Zhong

**Content**

**[1. Consultancy Profile 1](#_Toc6647)**

[1.1 Market of Consultancy 1](#_Toc30494)

[1.2 Target Clients 1](#_Toc10536)

[1.3 Focus of Consultancy 2](#_Toc9919)

**[2. Data Wrangling 3](#_Toc25161)**

[2.1 Datasource and Description 3](#_Toc1130)

[2.2 Data Quality 4](#_Toc2749)

[2.2.1 Initial Data Quality 4](#_Toc12496)

[2.2.2 Data Preparation Process 5](#_Toc21983)

[2.2.3 Data Quality Assurance 6](#_Toc6542)

**[3. Data Visualization 7](#_Toc3035)**

**[4. Model Building 27](#_Toc15128)**

**[5. Recommendation 31](#_Toc8297)**

**[6. Additional Data Sources 33](#_Toc18188)**

**[Reference 34](#_Toc76)**

# Consultancy Profile

## 1.1 Market of Consultancy

With the strong demand of the gaming console industry, the whole industry has been growing in the past five years. The total global gaming console market has reached a value of 10.30 billion and it's expected to grow to 71.73 billion by 2027. It is worthwhile for us to analyze the sales of different types of video games and then try to summarize the factors of a successful video game. Based on the results, we will make a strategic plan for game-producing companies to produce new games to develop their markets.

## 1.2 Target Clients

Our target clients are video game-producing companies. We will do the research for our clients to give them strategic insight into the industry through market research and industry trend analysis. The video game producing companies can benefit from the research by focusing on developing the video games that are more liked by the customers and make reasonable strategic movement into the new market.

## 1.3 Focus of Consultancy

* **Collect requirements**

First, we will have an in-depth understanding of the characteristics of each company and then specify each game-producing company's requirements. We will hold multiple meetings to confirm our customized plan whether will meet our clients' needs well.

* **Collect and organize data**

Second, we will be based on our scheme to collect relevant data and make data wrangling processes. Having a good quality of data is of great importance for us to make further analyses and ensure our recommendations are rational and credible.

* **Visualize data**

Third, based on the cleaned data, we will analyze it and use Tableau and R to visualize it so that we can offer insightful and easy-understanding reports for our clients.

In this case, we will make a strategic plan for game-producing companies to produce new games. Here are general questions we will put up with to further our analysis:

I. Which year has the largest number of video games been released?

II. Which company is the most competitive for which type of game？

III. What's the maximum, minimum, and average of video games’ Global Sales?

IV. Which console had the most games and how well they performed?

V. What’s the distribution of Global Sales for each platform?

VI. Which game has the highest sales?

VII. Which publisher was the most popular based on the total number of released games?

VIII. For different groups, which types of video games are the most popular?

IX. Is global sales related to user scores or critic scores?

# 2. Data Wrangling

## 2.1 Datasource and Description

This dataset comes from the Kaggle website and was generated by a scrape of vgchartz.com. It describes detailed information about video games and their developers. The following is an explanation of some fields.

|  |  |
| --- | --- |
| Fields | Description |
| Rank | Video games ranked by sales |
| Name | The games name |
| Year | Year of the release of the game |
| Platform | The platform of the release of the game |
| Genre | Genre of games |
| Publisher | Publisher of games |
| ESRB\_Rating | ESRB Rating of the game ( E means everyone; E10+ means everyone 10+; T means Teenager; M means Mature; EC means Early Childhood; RP means Rating Pending) |
| Global Sales | Total worldwide sales (in millions) |
| Developer | Developer of games |
| Critic\_Score | Critic Score of the game out of 10 |
| User\_Score | User Score of the game out of 10 |
| Country | Country of publisher |
| Year\_Foundation | Year of the foundation of publisher |

## 2.2 Data Quality

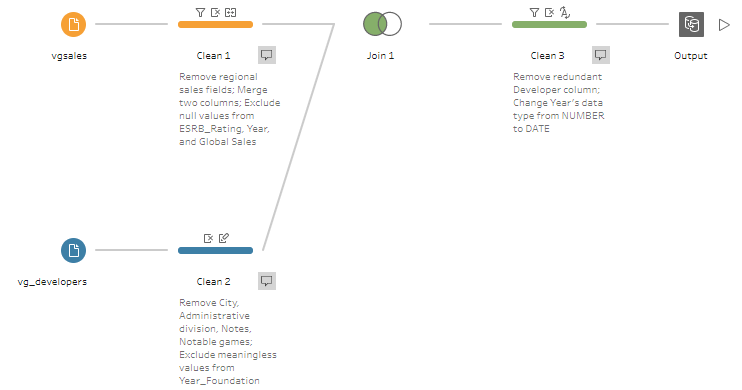
### 2.2.1 Initial Data Quality

We have two datasets: the first one is about video games sales called vgsales and the other is the description of developers called vg\_developers. The first dataset vgsales has a quite good data quality that the critical fields (such as Name, Year, Platform, Genre, Publisher, Developer) are completed and most of the data reflect the real-world things. But there are some sales columns including inconsistent data (such as NA\_Sales, PAL\_Sales, JP\_Sales, and Other\_Sales) and Total\_Shipped and Global\_Sales are redundant columns describing the video games' global sales. Then the second dataset vg\_developers includes detailed information about developers, but there are many missing data in Notable games, Notes, City, Administrative division columns.

### 2.2.2 Data Preparation Process

In order to ensure high data quality in our report, we modify our dataset as follows.

* In vgsales, we decided to drop NA\_Sales, PAL\_Sales, JP\_Sales, and Other\_Sales columns. Besides, we combined Total\_Shipped and Global\_Sales into one field called Global\_Sales and removed these two old columns. Since the mean imputation method does not work here because every video game and developer has its properties, we decided to exclude the null values in Year, ESRB\_Rating, and Global\_Sales fields (See in Clean 1 Step).
* In vg\_developers, since Notable games, Notes, City, Administrative division columns are insignificant, we decide to drop these four fields. Also, we removed some data without realistic meaning, for example, some developers have two different years of establishment (See in Clean 2 Step).
* The data in Year and Year\_Foundation columns has a numeric datatype, which should be converted into date type (See in Clean 3 Step).



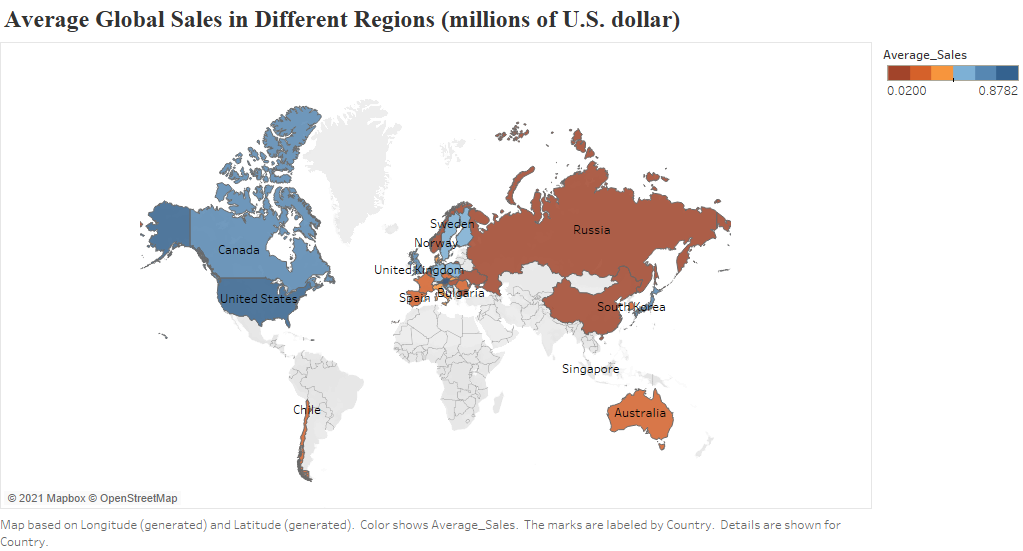
### 2.2.3 Data Quality Assurance

We use the "data quality dimensions" method created by DAMA UK to check our final data quality and find that our data is good enough to support our later analysis and ensure our results are more scientific and rational.

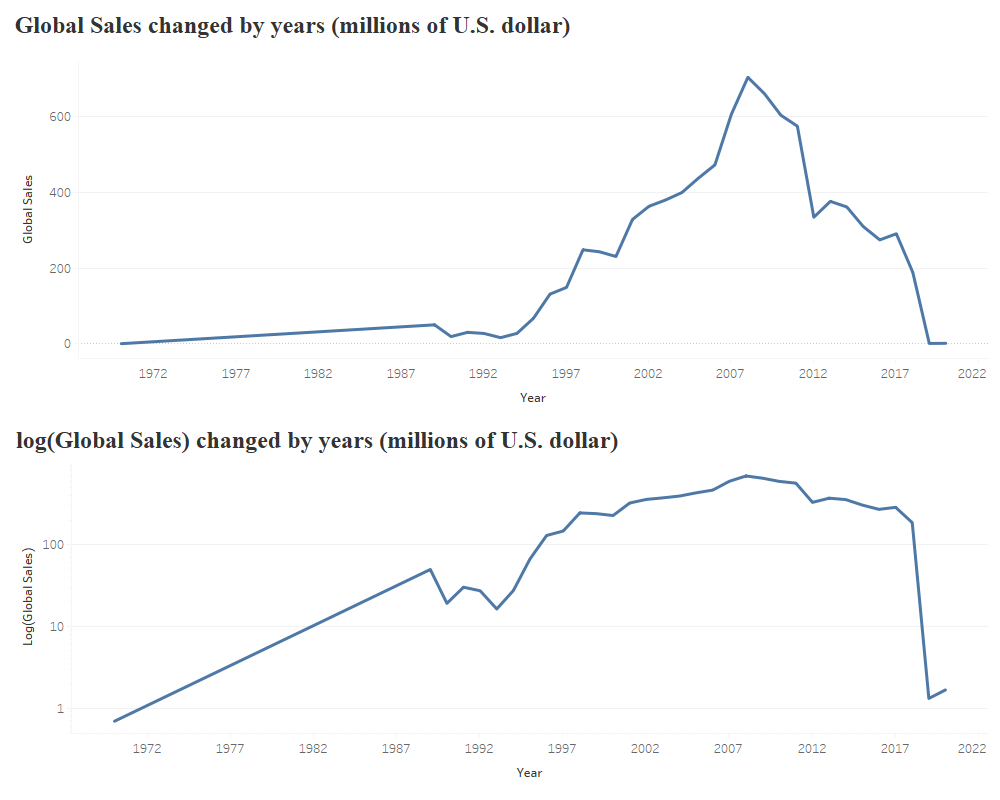
* Completeness: Critical data (such as Name, Year, Platform, Genre, Publisher, Developer) is completed.
* Uniqueness: In the critical fields mentioned above, every cell contains only one record.
* Timeliness: The released year of video games ranged from 1970 to 2020. Our data is still relatively new, and the results are also worthy of reference.
* Accuracy: After excluding some meaningless data, the data in the final dataset does reflect the real-world thing well. For example, some developers have two different years of establishment. Besides, the data type in Year and Year\_Foundation columns has been converted from numeric type to date type.
* Consistency: Data in critical fields is consistent, but there are some sales columns including inconsistent data (such as NA\_Sales, PAL\_Sales, JP\_Sales, and Other\_Sales) and Total\_Shipped and Global\_Sales are redundant columns describing the video games' global sales. We already drop those inconsistent fields, and the data in the Year column has the same pattern as /YYYY/.

# 3. Data Visualization

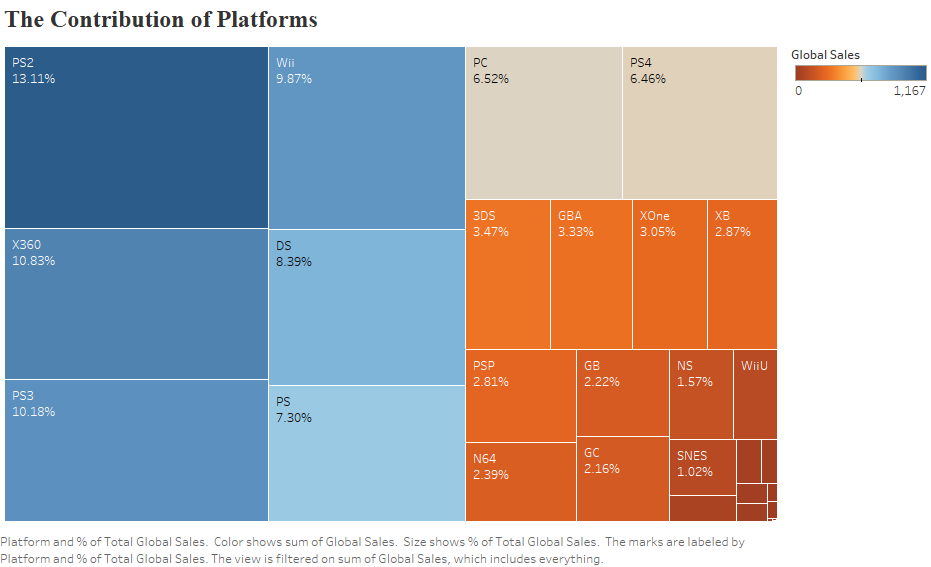
As a part of the research, we want to understand the industry-wide trend of video games. Initially, we want to understand where the maximum global sales were coming from and how the global sales figures have changed over time. Similarly, we also want to know the platform of the choice to create the video games and major developers. The popular genre of video games is also one of the interesting topics to dig into. Here are some of the findings of the research.



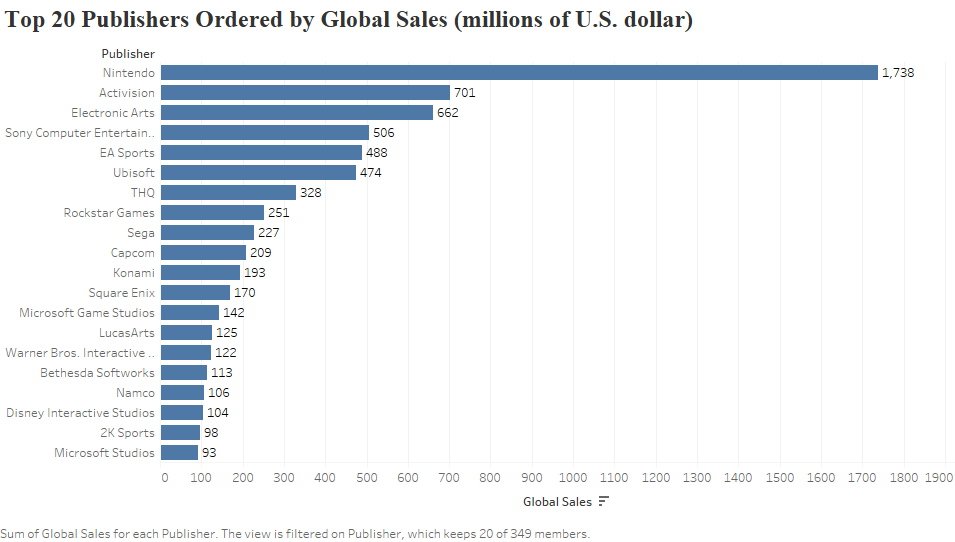
The North American region generates most of the sales of video games followed by Europe. Country-wise, the USA generates most of the sales followed by Canada, and Sweden, Norway, the UK, South Korea, and Japan. The video game industry has a lot of scope of growth because it is still not so popular in two of the largest countries of the world: India and China. The sales generated from these countries can be massive and provide lucrative profit opportunities.



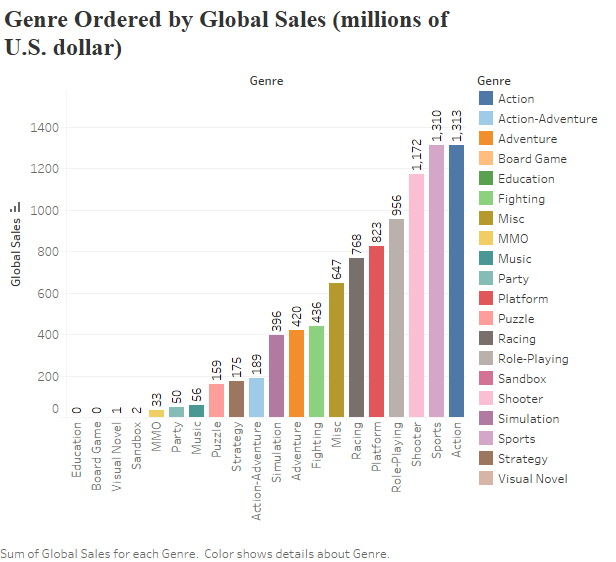
The video game industry started to become popular with the beginning of the dot com bubble. Its sale continued to increase until 2008 when the recession hit the US economy. Since the major producers and developers are from the US economy, the video game industry has also witnessed a massive drop in growth since then.



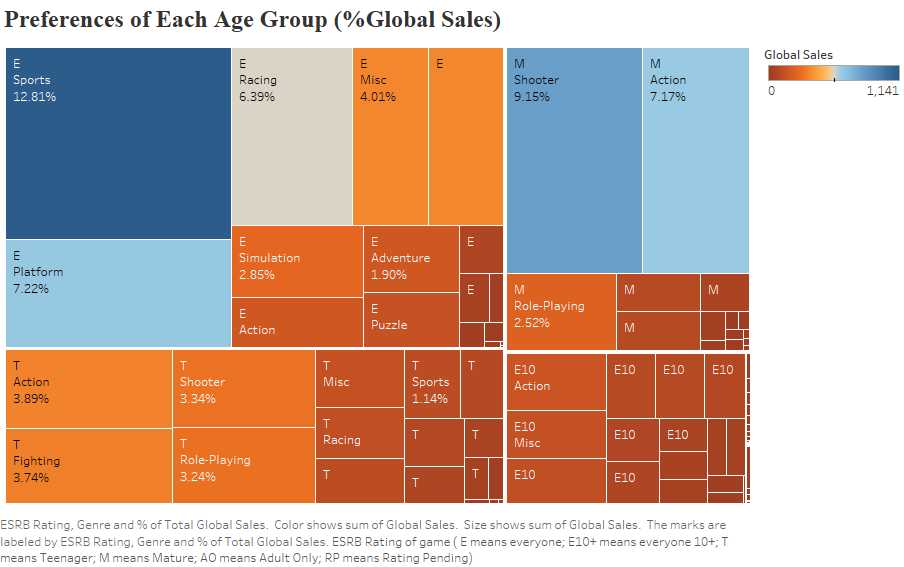
PS2 remains one of the favorite platforms for the developers of video games. While there are other alternatives that are growing like X360. PS3, Wii, and DS, PS2 dominates the market with market leader position. The primary reason for its market leader position can be attributed to the fact that it is user-friendly and supports creativity in a lot of different ways.

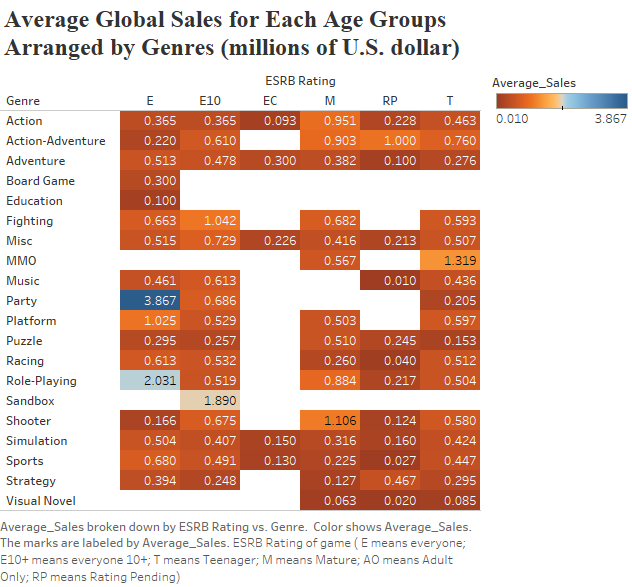


Nintendo has been retaining the position of market leader for a long time now followed by Activision and Electronics Arts.

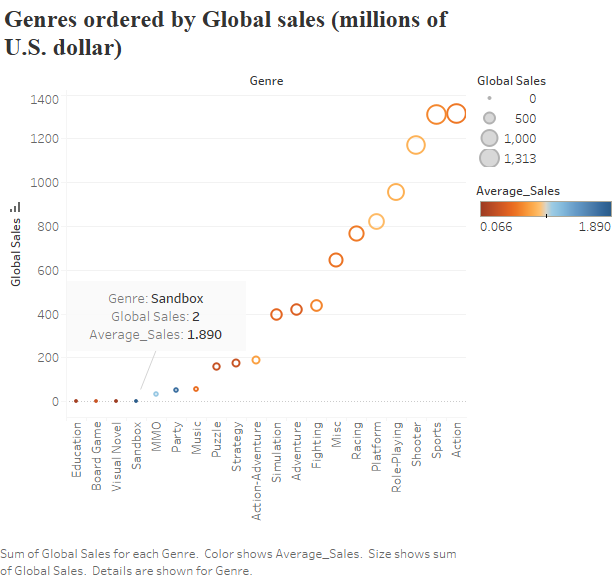


The figures show the sales figure categorized by the genre of video games. The action genre is more popular followed by sports, shooter, role-playing, etc. Racing, platform, role-playing are also popular genres of video games amongst the users.

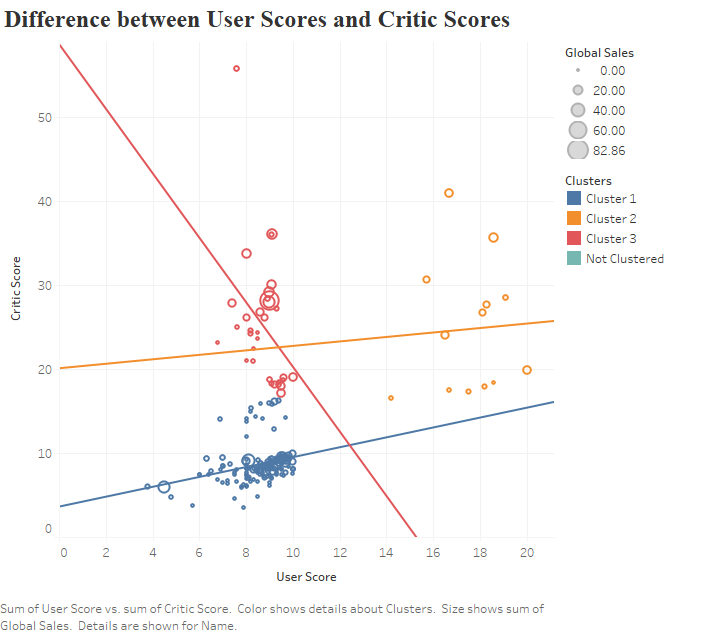
We can see the global sales of video games for E(everyone) was the highest, with Sports sales accounting for about 13% of the total and Platform sales about 7%. Then that for T(Teenager) was the second-highest, with Action and Fighting sales taking up about 4% respectively. The third-largest age group was for M(mature), with Shooter occupying 9% and Action 7%. Other groups like E10(Everyone 10+), EC(Early Childhood), and RP(Rating Pending) just accounted for a small proportion of total global sales. It's obvious that different age groups have various tastes, especially for age-restricted games. For game-producing companies, it is very important to identify the main audience of our games and conduct targeted marketing based on their preferences.



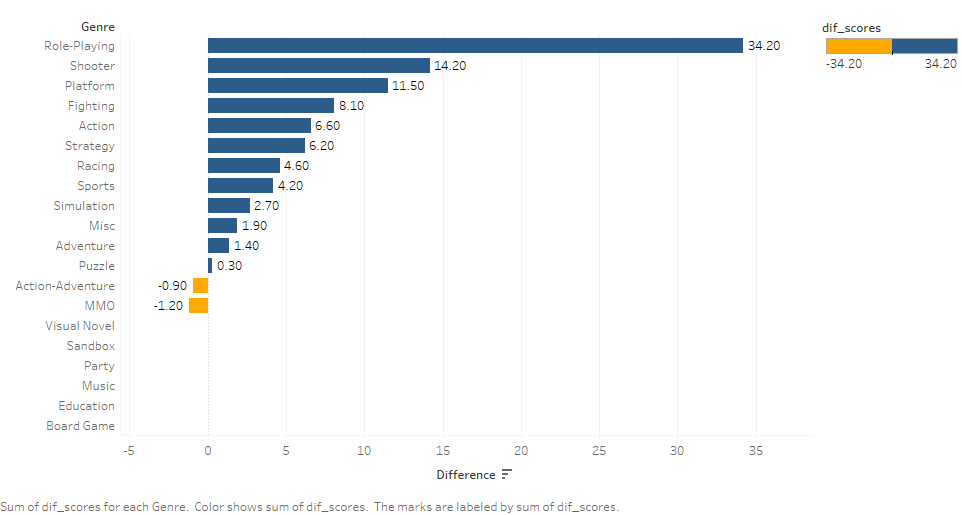
This is a highlighted table to list the average sales that different age groups spent on different genres of video games. A darker blue cell means higher average sales and the meaning of a darker orange cell is reversed. We can see only Party games for E(everyone) with the darker blue marks, followed by Role-Playing games for E(everyone) with grey marks and Sandbox for E10(Everyone 10+) with light yellow marks. Others like T(Teenager), M(mature), EC(Early Childhood), and RP(Rating Pending) are almost orange or darker orange. In contrast with the above tree chart, although the global sales of Party games and Role-Playing games for everyone did not contribute much to the total sales, their average sales were outstanding and attractive.



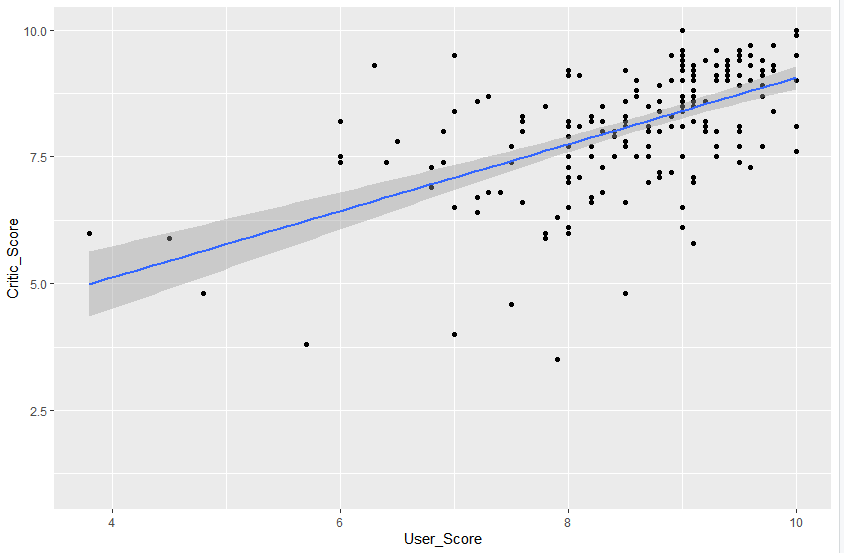
Sports, Action, Shooter, Role-Playing genre has the highest sales while education, board game, Visual Novel, Sandbox has the least sales. However, it is an interesting find that the genre with the highest global sales has the least average sales within the given genre. This might be because the action genre is very popular and has a lot of games within its space pulling its average down. In the case of the genre like education which has very few games within its space, the average is pushed upward.



The chart tells about the relationship between User\_Scores and Critic\_Scores. Here, we divided these games into three clusters: Bluepoints stand for Cluster 1, indicating that lower User Scores accompanied by lower Critic Scores; Orange points stand for Cluster 2, indicating that higher User Scores accompanied by higher Critic Scores; Red points stand for Cluster 3, indicating that lower User Scores accompanied by higher Critic Scores. Generally, we might think these two factors have a positive relationship, but we can see in chart 1 that there is also a negative trend line between the two and some of the points have a bigger bubble size indicating higher global sales. So the assumption does not stand and there is no exact relationship between User\_Scores and Critic\_Scores in general.

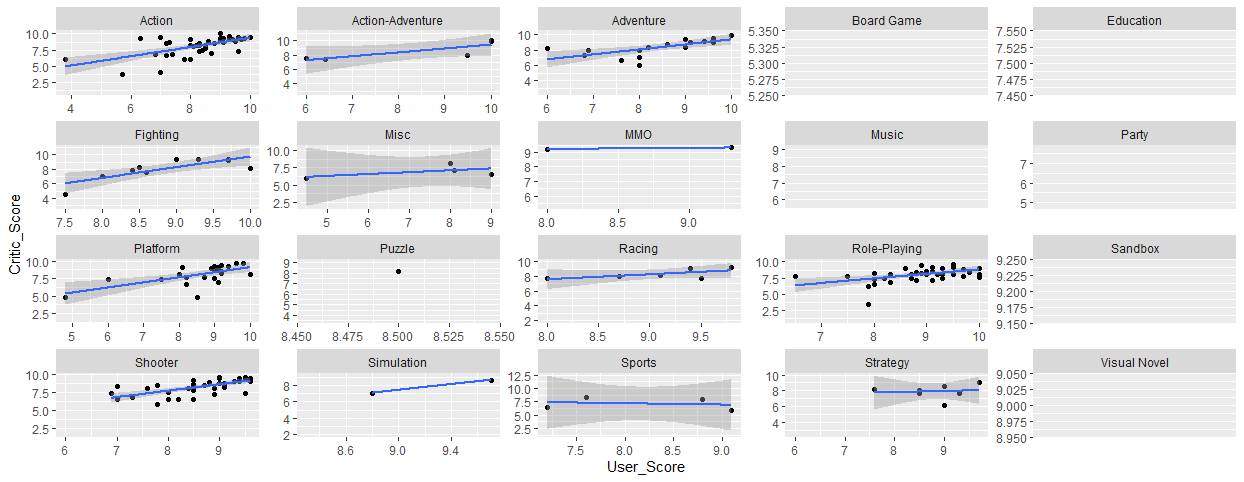
This chart tells about the difference between User\_Scores and Critic\_Scores and it also proves that there is a difference between User\_Scores and Critic\_Scores by various extents. For example, for Role-Playing, Shooter, and Platform, the difference between User\_Scores and Critic\_Scores exceeds 10 million. But that for other genres is less than 10 million, especially for Misc, Adventure, Action-Adventure, and MMO less than 2 million. When talking about the differences, many are positive indicating that User\_Scores surpasses Critic\_Scores, but two of them, Action-Adventure and MMO, are negative, which means User\_Scores are lower than Critic\_Scores.

**The correlation between user\_scores and critic\_scores**



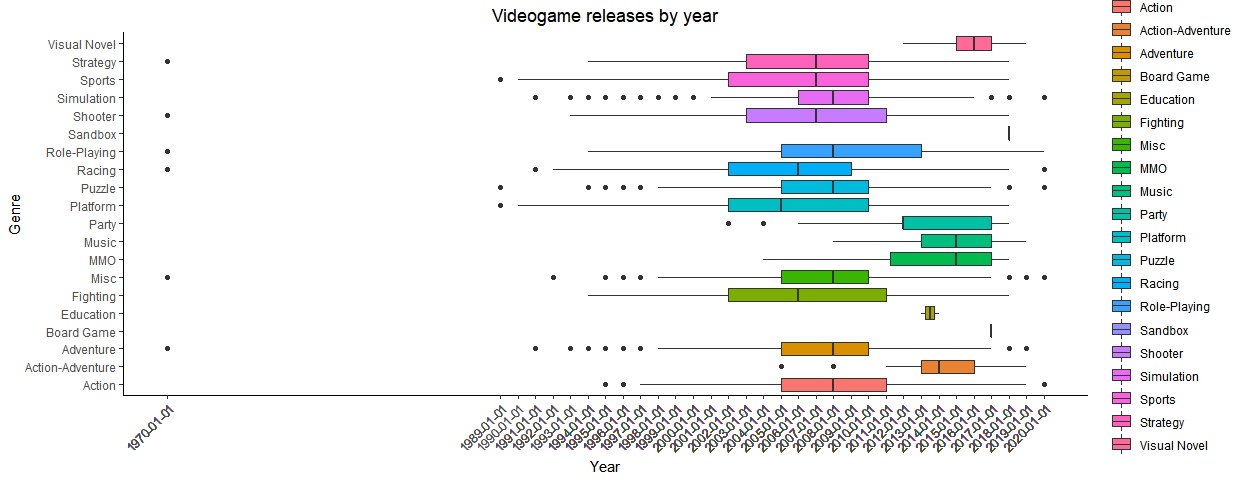
The figure shows that generally user\_scores and critic\_scores are in a positive relationship, which means higher user\_scores will be accompanied by higher critic\_scores. But we can also see there are some outliers that higher user\_scores are accompanied by lower critic\_scores, indicating that users and critics have different opinions on the same games. Later, we will dig out why there are outliers and whether there are patterns behind them.

**The correlation between user\_scores and critic\_scores grouped by genres**



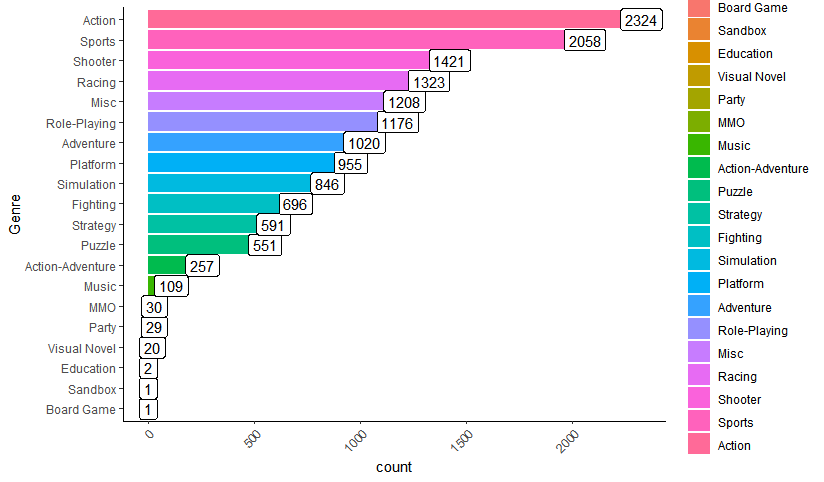
The figure is used to answer the questions raised in the last chart. We can see user\_scores and critic\_scores in Action, Fighting, Platform, Shooter, Adventure, and Role-Playing have an obviously positive relationship, indicating that user\_scores and critic\_scores will be approximate when evaluating the same video game in these genres. Besides, user\_scores and critic\_scores in Action-Adventure, Misc, and Simulation have a slight positive correlation and the supported data is less. However, user\_scores and critic\_scores in MMO, Racing, Sports, and Strategy, have a quite flat relationship and in the left genres have no pattern. For these genres with unclear correlations, we can not conclude there is a correlation between user\_scores and critic\_scores.

**The distribution of the released number of video games grouped by genres**



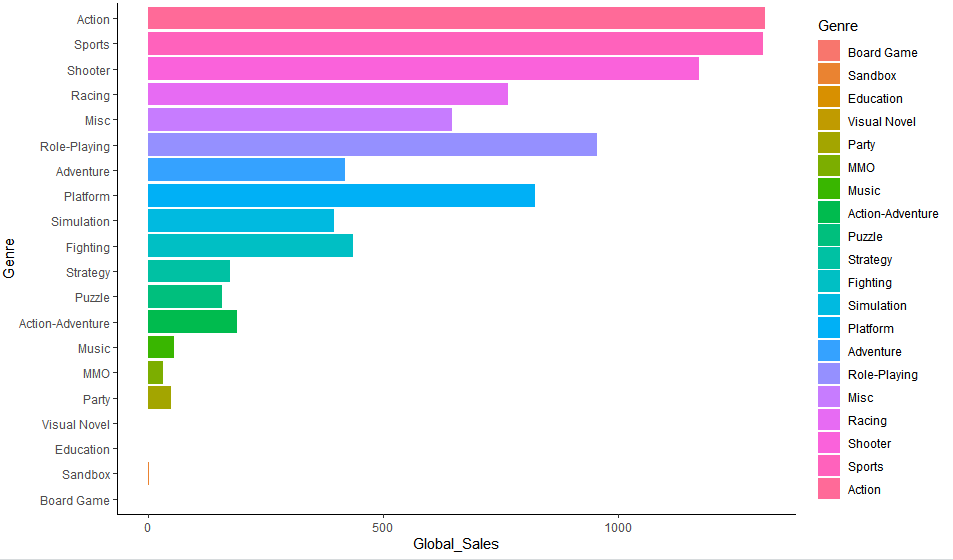
The figure shows the box plot diagram of the number of video games released per year grouped by genres. While most of the genres were popular during 2004-2010, there are some genres that became popular in the mid-2010s, such as Visual Novel, Sandbox, Party, Music, MMO, Education, Board Game, and Action - Adventure. Among these later popular games, we can see Sandbox and Board Game are the latest genres, enjoying popularity in the same time about 2017-2018.

**Genre listed by the released number of video games**



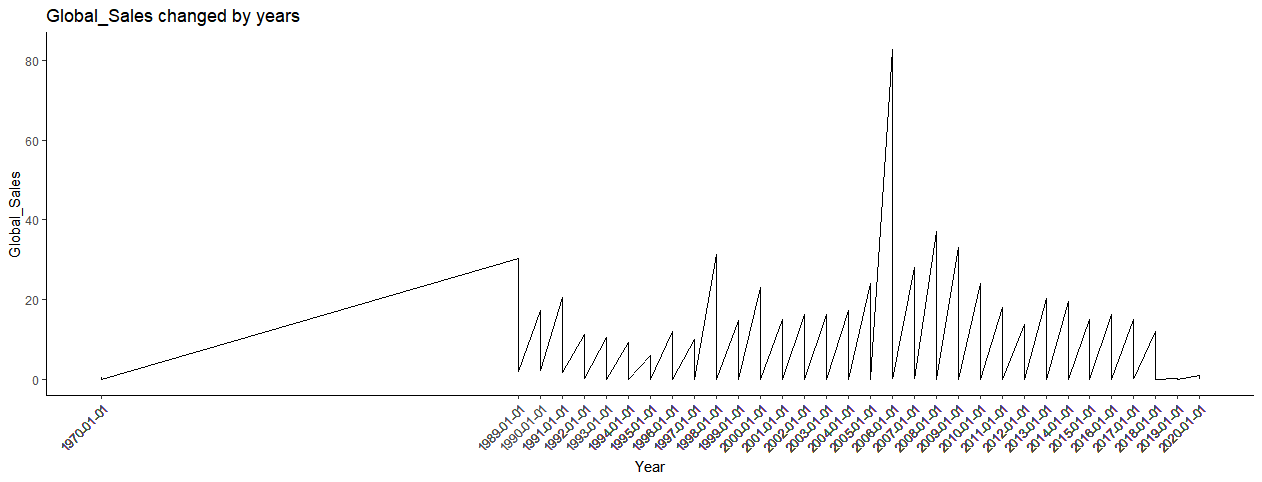
The figure shows the number of video games released grouped by genre. It can be seen that action and sports are the most popular genres with more than 2000 released games.

**Genre ordered by Global Sales**



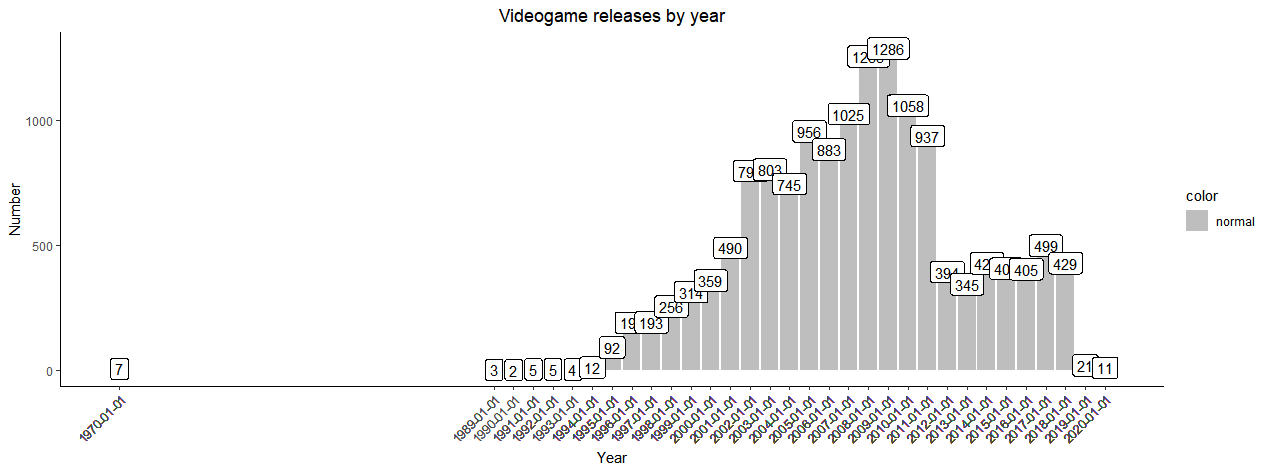
The figure shows that Action, Sports, and Shooter are the top 3 genres with the highest global sales. This ranking is the same as that in Figure 2. But we can see Role-Playing and Platform ranked fourth and fifth in the Global Sales ranking but sixth and eighth in the released number ranking respectively. Therefore, we can pay more attention to Role-Playing and Platform since these two genres have quite high global sales but the less released number of video games.

**Global Sales changed by Years**



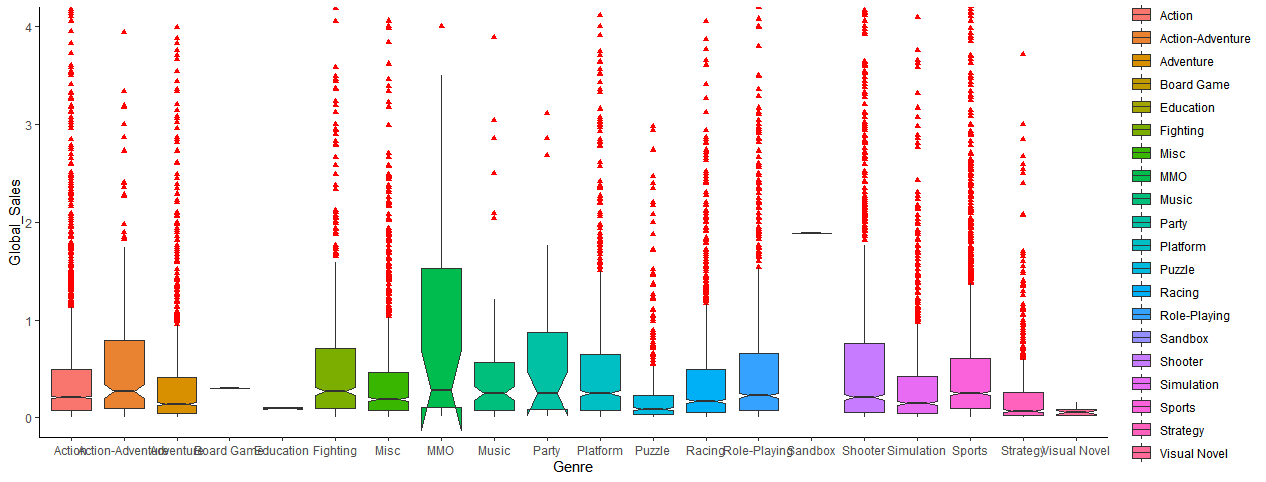
The figure shows the global sales of video games each year. It can be seen that the global sales ranked the first place in 2005, but experienced a downward trend since then. To be more specific, we can see global sales in the period between 2013 and 2016 increased slightly and then dropped back to the 2012 level.

**The number of released games for each year**



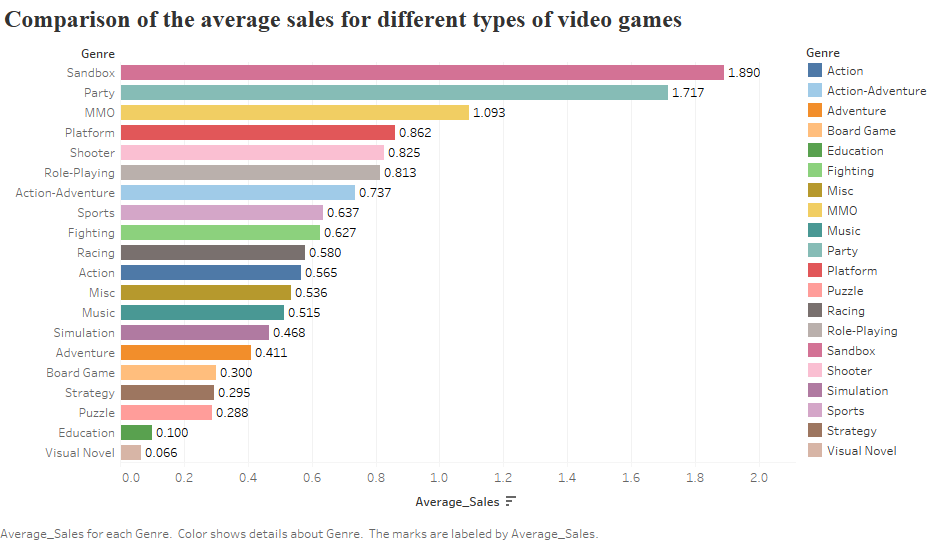
The figure shows the number of video games released each year. It can be seen that the sale of video games increased from the beginning of the internet era in the late 90s to 2008. The release number of video games has dramatically decreased post-recession, which almost matches the trend. But we can see the year 2009 had the highest number of released games but the year 2005 had the highest global sales. So there is a time delay, correctly reflecting that the more games a publisher releases, the market volume will not be better.

**The distribution of Global Sales grouped by Genres with outliers**



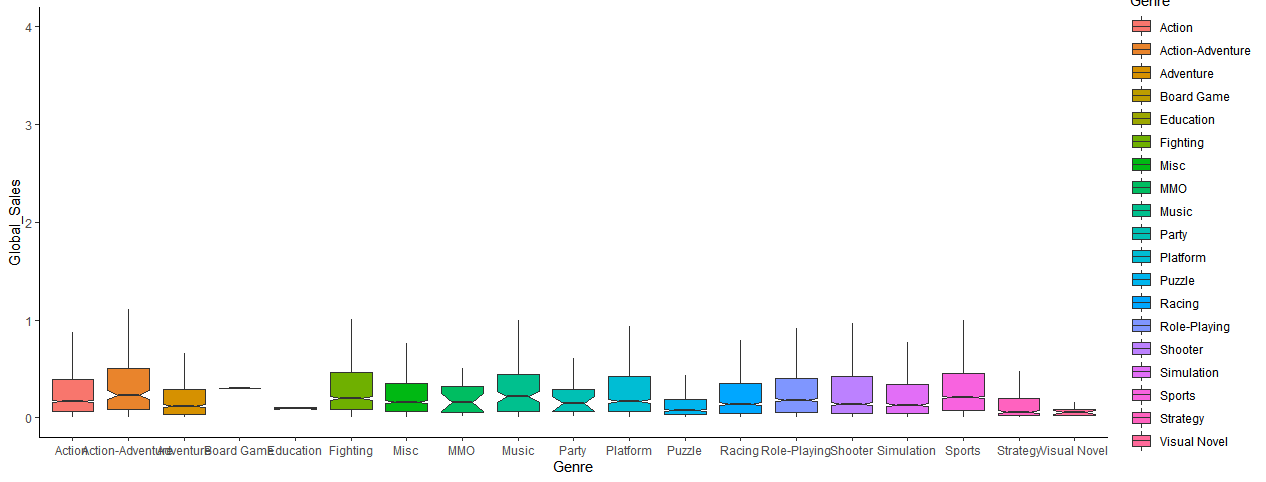
The figure describes the distribution of global sales grouped by genres, and the red triangle stands for outliers. There are two charts in Figure 2: the first one is the whole picture and the second one is the partial picture. We can see the distribution of the global sales of Board games, Education, Sandbox, and Visual Novel is more concentrated than other types of video games. Because Board Game and Sandbox are the latest types, with only one game published. Visual Novel enjoys great popularity with games continuously published during 2014-2019. On the contrary, Education is quite old-fashioned, since it only published two games in 2013-2014. For other types of video games, we can see Sports games have lots of outliers with a special one called Wii Sports earning 82.86 million dollars.

**Comparison of the average sales for different types of video games (millions of U.S. dollars)**



We can see that although Visual Novel is quite popular with games published every year since 2014, its average sales are the lowest with only 0.066 million dollars, followed by Education with the second-lowest average sales 0.1 million dollars. When talking about the latest games, Sandbox has the highest average sales of 1.89 million dollars, but Board games have a quite low average sales of 0.3 million dollars. Combined with the last Figure, we can see Sandbox has a great potential market with few competitors but many fans. Also, Party games are worth noticing, since it has the second-highest average sales and a quite concentrated distribution of sales.

**The distribution of Global Sales grouped by Genres without outliers**



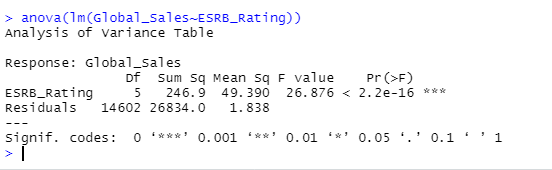
Compared with the chart with outliers, we can clearly see that the distribution of global sales without outliers for all types of games is more concentrated than before. Except for Board Game, Education, Visual Novel, and Sandbox (being excluded automatically because of small figures), the distribution of other types of games is similar to each other. If we don't consider the outliers, Genre might not be an important factor for building a successful game.

# 4. Model Building

Here we try to build a model to figure out how ESRB Rating and Critic Scores affect Global Sales:

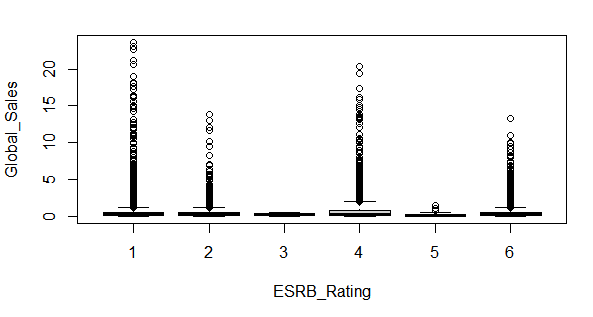
*Global\_Sales = β0 + β1 ESRB\_Rating + β2 Critic\_Scores + σ*

Where Global\_Sales is the global sales for every video game; ESRB\_Rating is a categorical variable, with E = 1, E10 = 2, EC = 3, M = 4, RP = 5, T = 6. And E means everyone, E10+ means everyone 10+, T means Teenager, M means Mature, EC means Early Childhood, and RP means Rating Pending; Critic Scores are the scores given by critics for particular games.

* **ANOVA Analysis** 

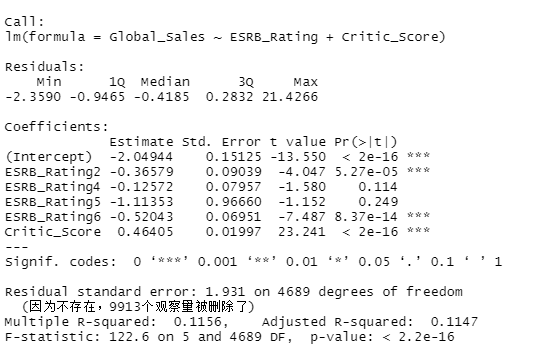
We conduct an ANOVA analysis to specify independence between continuous and categorical variables. The null hypothesis is that there is the same mean for all groups and the alternative hypothesis is that at least one mean is different. From the result, we can see a p-value < 0.05, so we reject the null, indicating that at least one group's mean is statistically different.

**The relationship between ESRB\_Rating and Global\_Sales**



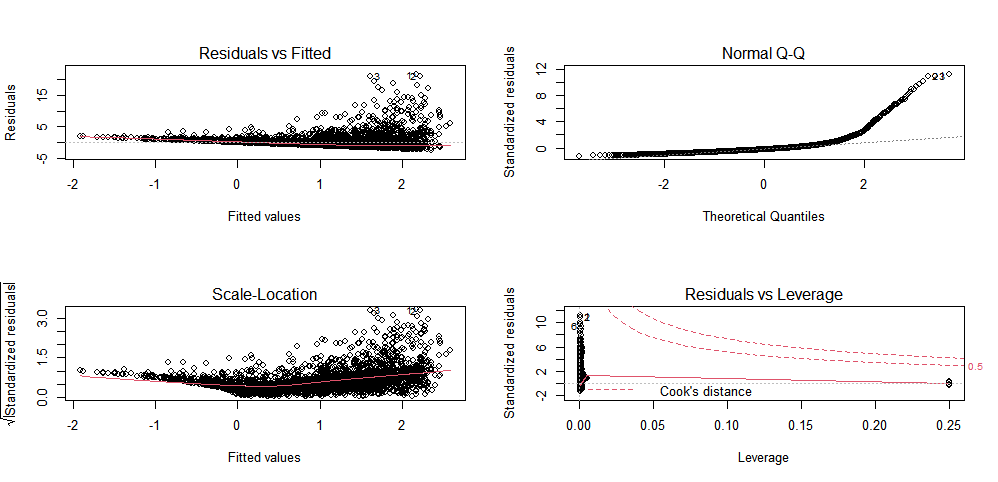
The figure proves what we conclude from ANOVA analysis that at least one group's mean is statistically different.

* **Regression Analysis**



We can see Critic Score is statistically significant, indicating that when everything being equal, 1% increase in Critic Scores will lead to a 0.46405% increase in Global Sales. When talking about ESRB Rating, we can see only E10 (ESRB\_Rating2) and T (ESRB\_Rating6) are statistically significant, indicating that global sales from video games for E10 (Everyone 10+) is 0.36579% lower than those for E (everyone), and sales from games for T (Teenager) is 0.52043% lower than those for E (everyone). But for other age groups, it's hard to conclude since their p-value is greater than 0.05, which means they are statistically insignificant.

**Residual plots for model**



We can see there are too many outliers in the dataset, which leads to a poor fit of our model, and the trend line deviates from the standard line. In the next step, we try to exclude these outliers to make our model a better fit.

# 5. Recommendation

* **In terms of Region**, the North American region generates most of the sales of video games, followed by Europe. It is still not so popular in two of the largest countries of the world: India and China, which will provide lucrative profit opportunities.
* **In terms of Platform**, PS2 dominates the market, while there are other alternatives that are growing like X360. PS3, Wii, and DS.
* **In terms of Genres**, Actions, Sports, and Shooters are the top 3 genres with the highest global sales and the largest released number of video games. Besides, Sandbox, as one of the latest genres, is a potential market worth exploring since it only published one game with the highest average global sales. And Party games need to be noticed, as they have the second-highest average global sales and this market is more mature and stable than Sandbox. Surprisingly, if we don't consider the outliers and the latest types of games, Genre might not be an important factor for building a successful game as many types of games have a similar distribution of global sales.
* **In terms of User Scores and Critic Scores**, we find users and critics have different opinions on the same games. Generally, user\_scores and critic\_scores in Action, Fighting, Platform, Shooter, Adventure, and Role-Playing have an obviously positive relationship but the rest of the genres have an unclear correlation. What's more, higher Critic Scores will lead to higher global sales, indicating that it's useful for game-producing companies to hire critics to spread good comments to the games, which will attract users to play.
* **In terms of Age Groups**, video games for E (everyone) will have higher global sales than others. It is best for a game-production company to make video games suitable for a wider range of users. Also, different age groups have various tastes, especially for age-restricted games.

# 6. Additional Data Sources

Having the unit price of each game is very helpful for us to develop marketing strategies. In this way, we can analyze the sensitivity of each age group to game prices. At the same time, when promoting products, customized marketing is carried out according to the characteristics of each group to maximize the company's profits. Moreover, it would also help to know if the people prefer to play new levels of the same video games over time or switch to a different video game and switching frequency.

# Reference

[1] "Video Games Sales 2019". 2021. *Kaggle.Com*. <https://www.kaggle.com/ashaheedq/video-games-sales-2019>.

[2] "Videogames-Companies-Regions". 2021. *Kaggle.Com*. <https://www.kaggle.com/andreshg/videogamescompaniesregions>.