Yuvraj Sreepathi
ys3mnq@virginia.edu
11/25/2020
Final Project: Compiled Report – Analysis of Top Video Games of All Time

Abstract

With the new generation moving into an era of digital information and the need for massive amounts of data to be stored, the entertainment industry is following a similar pattern. Specifically, the video game industry, which is largely unknown to a majority of the population, is experiencing rapid growth with estimates of around $153.9 billion in revenue. This industry is also expected 9.3% growth annually. To survive in this market, companies must continuously produce best-selling games to maximize their profits. This study analyzed the factors that contributed to a game's overall success and what steps a company should take in producing its next game. The initial purpose of the study was to find a correlation between the score's games received from certified critics and the number of global sales the game had. After this correlation became evident through a series of visualizations and confidence tests, the next step was to determine whether games with a specific ESRB Rating tended to perform better than one another. The two most popular ratings that were studied were "Everyone" and "Teen." After performing an independent two-sample hypothesis test, I found that there was not enough evidence to prove that one Rating provided more sales or had higher critic scores than another. This would go to show how competitive this industry can be and the need for companies to constantly bring forth new ideas to stay ahead of their competitors. The main conclusion of this study was that companies should look into factors that can maximize the critic score their games receive. Doing so, would in turn maximize the video game's overall profit.

Introduction

To someone not familiar with the gaming industry, this study could come off as irrelevant and childish. However, the entertainment industry which includes video games accounts for about 5% of the US Economy. Studying a dataset that compares various statistics and measurements of some of the most well-known titles gives perspective on how products can be broken down into so many metrics. While it may not have the most practical applications for analyzing this dataset, getting a greater understanding of this multi-billion-dollar entertainment industry can give companies ideas on general trends among their consumers. This would eventually lead to even more growth in the industry and increased satisfaction of customers. To get a better understanding of this, I needed a reliable dataset to work with that contained the necessary variables of interest. The dataset that was used was called "Video Game Sales 2019" and was retrieved from Kaggle. It was posted by Abdulshaheed Alqunber on December 4th, 2019. The file contains 16 fields with a mix of categorical and quantitative variables. The variable in the first column is rank, which is based on the overall sales of the game. It ranges from 1 through 55,792 or the total number of entries in the dataset. The second field is the name of the game, which is self-explanatory as it is the distinguishing factor between the list of video games provided. The third variable is the platform. This essentially means what platform the game is available on. While some video games eventually became available on multiple platforms, they are listed separately if the games accumulated enough sales to become comparable to the most sold games of all time. The next variable is the genre, which groups video games into a category based on the contents within the video game. The ESRB Rating is another variable that distinguishes games based on the content of the game. It is a rating scale

that is accepted internationally to determine which age group the video game is suitable for. Likewise, two more categorical variables are the developer and publisher for a video game. While most times, these variables can overlap, sometimes the company that develops a video game can differ from the company that handles publishing copies of the game. From here, the variables discussed are now all quantitative. The first two being a critic score and user score. They are based on reviews of what people felt about the game distinguished by professional reviewers and ordinary customers. These ratings have nothing to do with the rank of the video game listed. The next statistic is the total shipped copies of the game. This value is what divided the rank of the video game among the dataset. The next four columns in the CSV are different amounts of sales recorded based on global sales, North American Sales, European Sales, Japanese Sales, and all other sales. These values are specifically made unique as it helps to distinguish where each game had more prevalence. The last variable is the year, which solely states when the game was released. With these variables in mind, we can accurately perform an analysis on this topic.

Content

The analysis was broken up into four distinct parts, which included wrangling, visualization, regression, and inferences. The entire process was meant to be successive where each part built upon the previous.

The data wrangling section was just a means of cleaning up the larger dataset and getting an idea of some of the key statistics of each of the variables. This phase of the study did not require much work as the data was fairly organized to start with. However, I first began with understanding each of the variables I was working with. Using R as a means of understanding the data, I used the “summary” method on all of the quantitative variables to get an idea of how the data was spread for each variable. This method was performed on critic score, user score, total shipped, and global sales. This method showed how skewed the data could be. Luckily, I could rely on measurements resistant to outliers such as the median and the different quartiles. Using this, I gained insight into all the key metrics that companies need to understand when it comes to releasing new games. I also came to realize that there are a lot of blank values within this dataset. This is understandable as it is not an official database regarding the information of the top video games of all time. However, it goes to show that the incompleteness of this dataset could potentially cause problems in the analysis of the data.

The next step was to subset the data to understand the key variables. The thought process behind this was based on different platforms that games were introduced on. I wanted to see how the leading consoles were performing compared to each other. Because these statistics are for all time, I could not just compare the most recent and currently popular consoles. However, I limited it to Global Sales for each categorical variable. By doing this I could see which platform had the most overall sales from games released on their console. Likewise, I could do the same with ESRB Rating, year released, and critic score. Essentially, the Global Sales were a reliable measure to show how the other categorical variables matched up to each other. Based on the breakdown of how the most popular games are related to these platforms, it will not only show which console is an all-time leader, but it will also prove how companies should plan future games. As someone who is deeply vested within the gaming industry, breakdowns like this are essential for getting an edge against competitors. Gaining the release rights for

upcoming games that are similar or related to the currently highest-ranked video games could do wonders for companies. By analyzing data sets such as this and seeing which company a series or brand of video game has partnered with in the past is insightful information. The breakdown of partnerships between consoles and all-time greatest video games would give information on the race/competition between the console war.

To complete the data wrangling portion of this study, I simply removed any data columns that were not being used. I removed most of the unused columns to bring the total down to 11 columns. From there, most of the data was already formatted in the order that I had intended it to be. However, the status column was a bit confusing, therefore I changed up the entries in that column to be more informative. The idea was to determine whether that specific video game entry was still available for purchase and if it could be used by the current consoles in the market. Finally, I renamed a few of the columns to be a bit more informative. Overall, this phase of the study provided the much-needed organization this dataset needed. *The tables produced can be found in the appendix.

The next step in the study was to create a more presentable approach to the wrangling done in the first part. This section focused on creating visuals of the tables that were derived in the wrangling phase. To do this, I looked at variables individually and then in comparison to one another.

The first variable that was represented was the ESRB Rating. For this, I used a bar graph to represent the number of games for each rating level. While there were many games without this rating in the dataset, the most distinguishable ESRB Rating was “E”. This shows that a majority of the games are made to be played by everyone and this gradually decreases as the target audience reaches a smaller pool. The second variable analyzed was the genre of the games produced. To do this, I created a word cloud of the genres whose size was determined by the count, and the color was determined by the ESRB Rating. In almost every color group, the most popular genre of game was Miscellaneous or Action/Adventure. This trend shows at which volume each genre of games is produced in comparison to one another. A similar pattern was followed in looking at the Platform each game was released on. I created another word cloud with the size based on the count and color based on the ESRB rating. Following the same trend as the Genre variable, I found that games released on the PC or PlayStation 2 were the most popular among all the platforms. The last categorical variable I looked at was the year the games were released. This data was easiest to represent in a bar graph as I was only dealing with the occurrences of the year in which a game was released. From the bar graph generated, I found games began to become more popular as time progressed and reached an all-time peak in 2009. More recently, there has been a downwards trend in the number of games released in the past ten years showing that games before then had performed better than they have now. After getting a breakdown of the categorical variables, I used a series of histograms and boxplots to analyze the numerical variables. The first two histograms I looked at were the Total Shipped and Global Sales of each game. These statistics were similar in the sense that they were highly right skewed with a peak very low. This shows that only a few games very highly successful in terms of sales and copies made, while the rest tended to accumulate among a very low success margin on sales. The next two variables were the critic score and the user scores the games given. The median for the critic score was about 7.5 and the median for the user score was about 8.5. both boxplots contained low outliers but no high outliers, showing that on average games tend to score fairly high ratings, except for a handful of games that do excessively worse in terms of review. The critic score variable seems more dependable as there are many more data entries compared to the user score.

The next visualization dealt with numerical-categorical and numeric-numeric relationships. For the numeric-categorical relationship, I created a boxplot of the critic scores based on the ESRB rating. While I did this for user scores as well, this data was more complete using the critic score. From this, I found that games that have ratings that appeal to a larger audience tended to have better critic scores compared to games that were limited to teens or adults. The numeric-numeric relationship compared the Critic Score to the global sales of a game. From this comparison, I found that games with a higher critic score tended to have more global sales than that of lower critic scores. This is a direct relationship.

The final visualization put it all together with three variables depicted in a graph. The bubble plot made from comparing ESRB Ratings, Critic Scores, and Global Sales provided the most meaningful insight. From this graph, I found that games with a higher ESRB Rating and high critic score tended to have the most global sales overall, and games with a low ESRB Rating and low critic score tended to have the smallest number of global sales. This shows there is a direct relationship between these three variables.

After creating useful visuals of the relevant data and variables, the next step was the regression phase. This dealt with finding relationships between variables in a more mathematical sense, specifically through correlation and linear regression. Throughout this phase, I compared various numeric variables to gain insightful data on how well these variables actually correlated. Compared to the previous two phases where we were simply representing the data set in different ways, providing regressions of key variables can help make accurate and intuitive predictions.

The four numeric variables I used were the Global Sales, Total Shipped, Critic Score, and User Score. The first linear regression I computed was between the Global Sales and the Critic Score. This correlation would provide the most accurate result as these data columns contained the most data. Using the Global Sales as the response variable and the Critic Score as the predictor variable, I found that this model has a median residual of -0.3103 which shows that resistant to outliers the data is not too far off the regression created. The second regression created had the critic score as the predictor variable again, but now with total copies shipped as the response variable. Because of overall having less data points, the linear regression created was less accurate with a median residual of -1.648. A similar pattern followed with the comparison between user score and the total shipped. It resulted in the highest median residual of -3.437. Because the Global Sales variable had the most amount of data entries, it would make sense that it would have a more accurate regression line. When comparing user score to the global sales variable, I found a median residual of -0.7906. These relatively low residuals among all the data points when comparing the two numeric variables shows that the difference between the line of best fit and the specific data points is fairly small. While the R-Squared values for these predictions are not that high, this could be due to the fact of so many outliers and missing data points that are shared across the numeric variables studied. With information on all the residuals and the R-Squared value it has shown that there is an obvious fairly strong, positive, linear correlation among the comparisons made.

I then created a logistic regression model. This simply meant to create a model based on multiple variables and their correlation. Because every numeric variable has many missing data entries, finding complete observations for a specific video game proved to be difficult. However, I was still able to create a correlation matrix between the Critic Score, User Score, and the Global Sales. From this, I used both the User and Critic Score as predictor variables for the Global Sales a video game had. Before

computing the model, I passed it through the MASS library using the stepAIC method. This made sure that both the User and Critic Score were accurate enough predictor variables. From this I found that the median residual was -0.4290, with an R-Squared value of 0.2557. Likewise, the slopes for the User Score and Critic Score were -0.1811 and 1.0836 respectively, showing a linear relationship. The standard error was also fairly low coming in at 0.2010 and 0.1698. From this I performed an interval prediction for global sales. I tested it for the value of 10 for both the user and critic score. The output was 3.756637, which is based in the millions. This is essentially the highest amount of possible sales a game would be expected to get based on ratings on 10 by the users and professional critics. Based on these values, it shows there is a stronger linear relationship between Global Sales and a combination of both User and Critic Scores. Using data like this, companies can go back and see which games had higher scores and view specific categorical variables among these games. Then, they can build upon them and make more overall revenue based on these past results.

After computing the regression models, we could move on to making inferences about the population itself based on the sample. By working with this sample data set with more than enough entries, we can get a clear picture of how successful games will be in terms of the entire population.

Because there was no valid conjecture that could be formulated with the data I had, I solely performed a one-sample confidence interval test and two-sample hypothesis test. The first test was a one-sample test to determine the mean critic score from the given sample. Based on the output, we are 95% confident that the true mean critic scores for all video games lie between 7.18 and 7.25. While this score may not be that high, it is commonly known that in terms of rating, a seven out of ten usually implies an average score. This would then show that if we considered outliers, the distribution of video games would be normal. Building upon the first confidence interval test, the next test dealt with two samples. By finding that the average critic score of games was above 7, I created a table based on the ESRB Rating and games that fell either above or below a critic score of 7. I also used the ratings of “E” and “T”, which stand for “Everyone” and “Teen” respectively. From this, I conducted a two-sample hypothesis test of independent proportions. My null hypothesis was to assume that the proportions between the two ratings would be equal. The alternative hypothesis was that the “E” rating would have a greater proportion of games with a critic score over 7. Using a 5% significance level, the computed p-value was 0.9527. Therefore, we don’t have enough statistical evidence to reject the null hypothesis. This is a logical computation as the computed proportions for the ratings were 0.57 and 0.59 respectively. They seem to be too close to make a definitive claim as to which rating is better/more successful.

Conclusions/Discussions/Limitations

After performing this study, there was a clear correlation between the Critic Scores games were receiving and the amount of sales they were making. Based on the various metrics shown and represented in terms of one another, it is clear that the market for video games remains dynamic and companies will need to consistently revolutionize they market and create games to remain competitive. Based on the study, companies should try to maximize their reviews from certified critics and see what factors play into a currently successful game. In addition, it doesn’t seem to matter which ESRB Rating a game is given as that does not act as a defining factor in how successful a game will be. This not only

ensures the gaming market will remain diversified, but it also allows companies to not be limited in the creative aspect of video game production. While my initial thoughts were incorrect where every metric of a video game would have an effect on its sales, throughout this research process I found the most significant factors and how the correlate to one another as well. This study provides the much-needed insight on the robust industry that could not be recognizable at a surface level view. Overall, it has shown the necessity of statistical analysis in a field so prominent in technology.

References

2017 Economic Impact Report. (2019, May 23). Retrieved November 28, 2020, from https://www.theesa.com/esa-research/2017-economic-impact-report

4 Ideas of Statistical Inference. (n.d.). Retrieved November 28, 2020, from http://www.bristol.ac.uk/medical-school/media/rms/red/4_ideas_of_statistical_inference.html

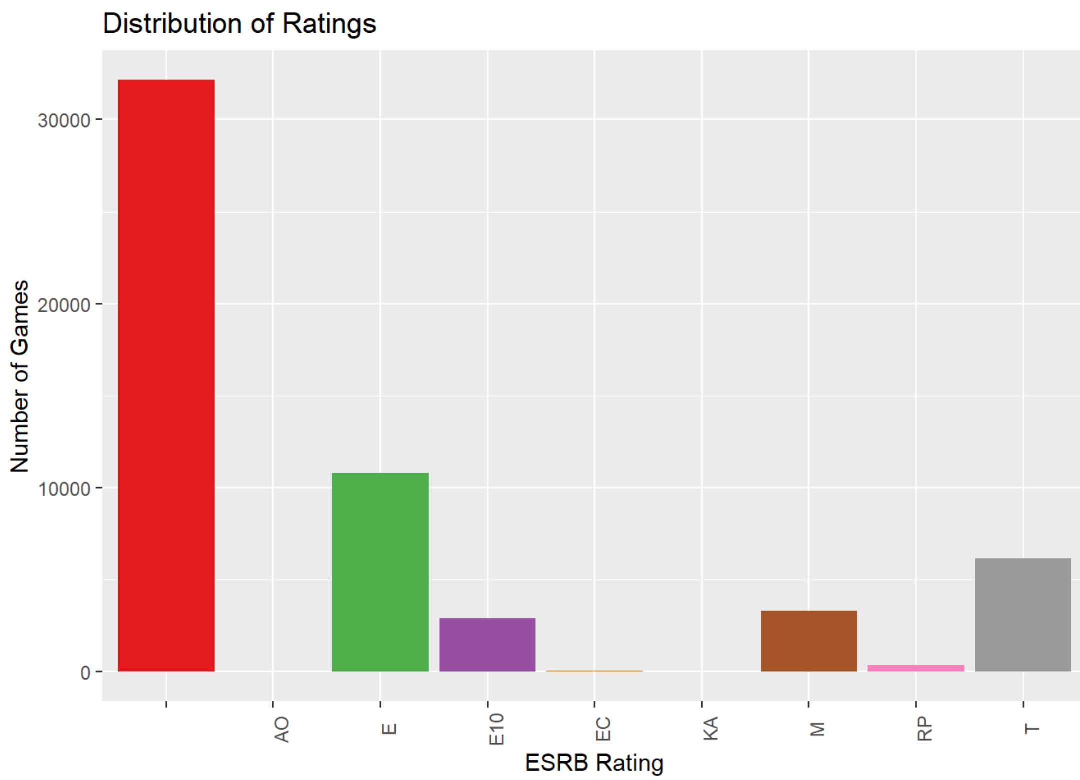
Alqunber, A. (2019, April 13). Video Games Sales 2019. Retrieved November 28, 2020, from https://www.kaggle.com/ashaheedq/video-games-sales-2019

Gordon, L. (2020, May 05). The many ways video game development impacts the climate crisis. Retrieved November 28, 2020, from https://www.theverge.com/2020/5/5/21243285/video-games-climate-crisis-impact-xbox-playstation-developers

Gough, C. (2020, September 26). Topic: Video Gaming Industry. Retrieved November 28, 2020, from https://www.statista.com/topics/868/video-games/

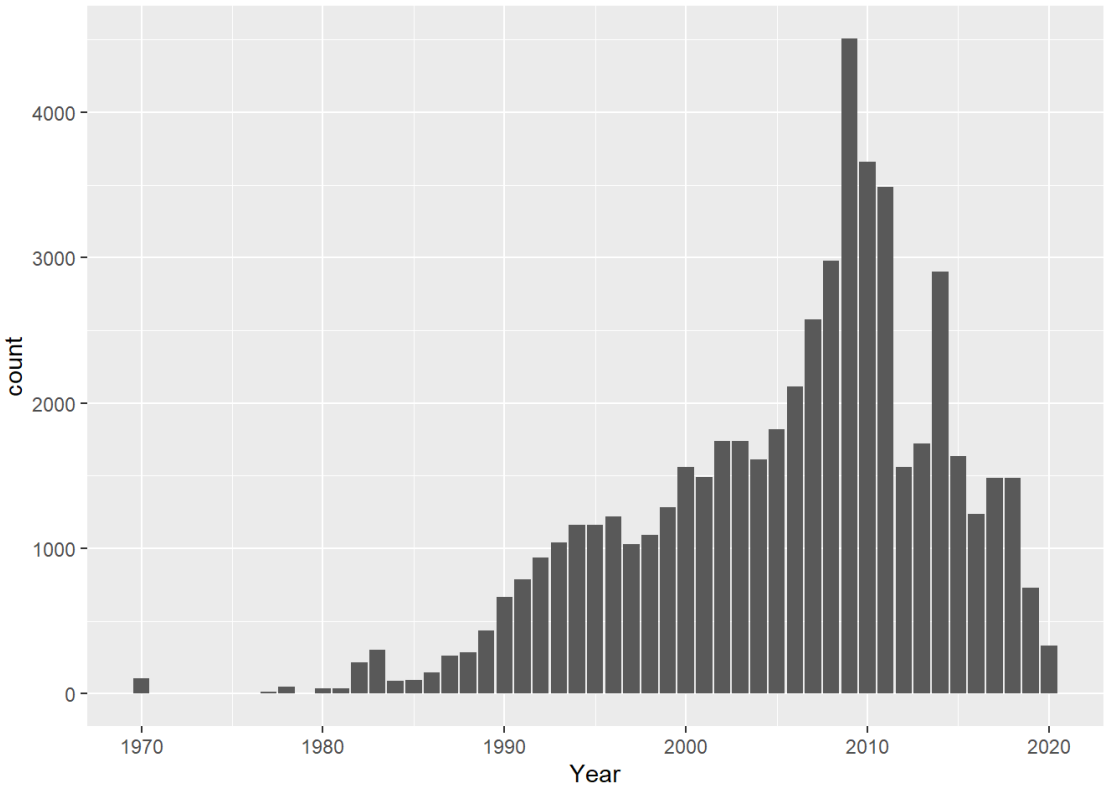
Will gaming keep growing when the lockdowns end? (n.d.). Retrieved November 28, 2020, from https://www2.deloitte.com/us/en/insights/industry/technology/video-game-industry-trends.html

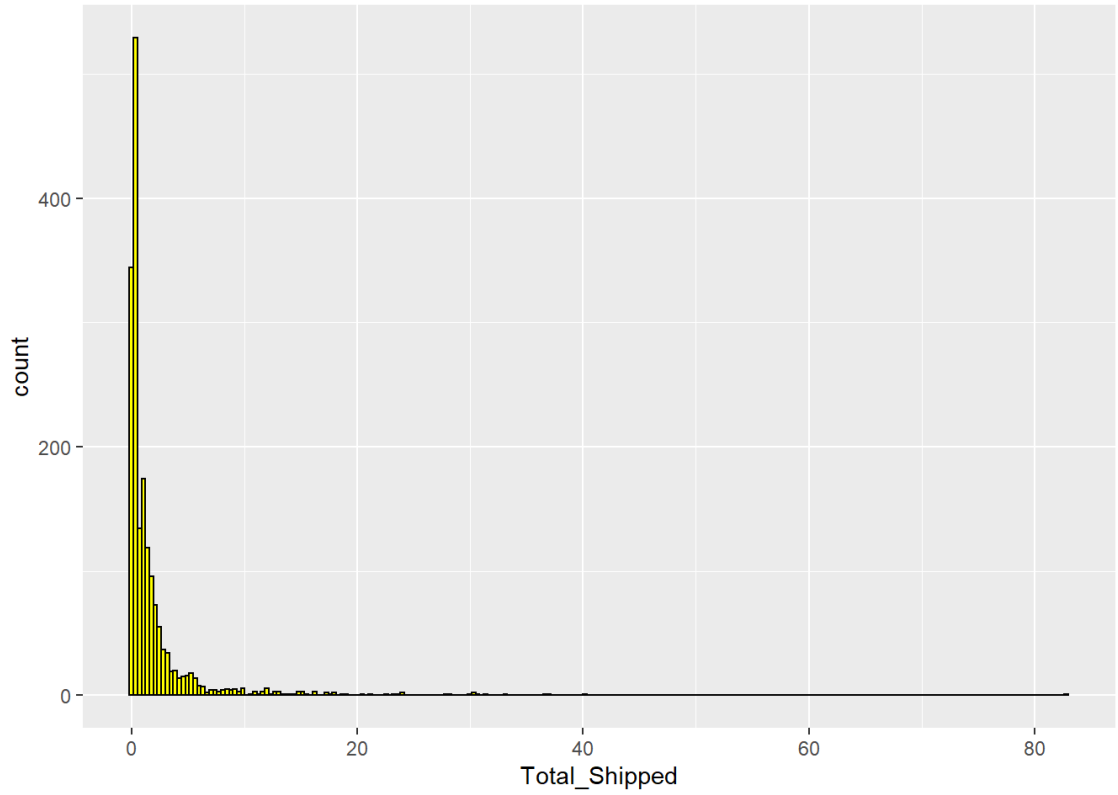
Appendix

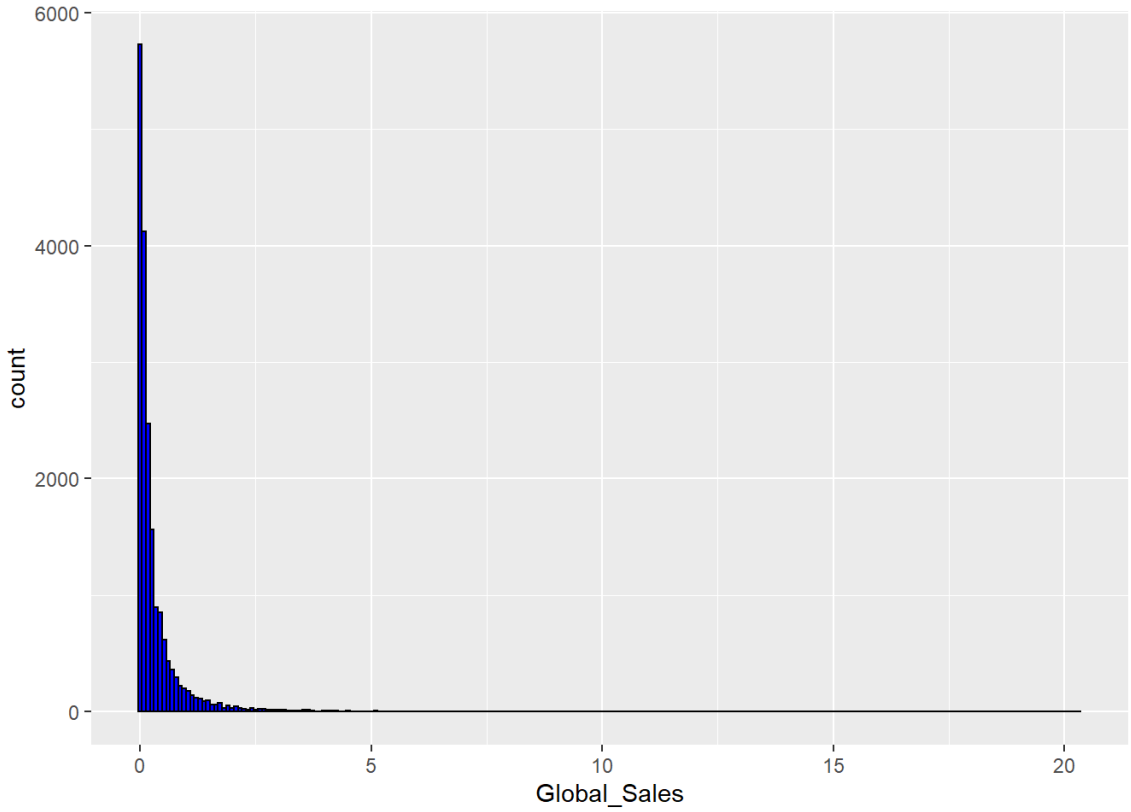

Distribution of Ratings
Number of Games
30000
20000
10000
0
AO
E
E10
EC
KA
M
RP
T
ESRB Rating

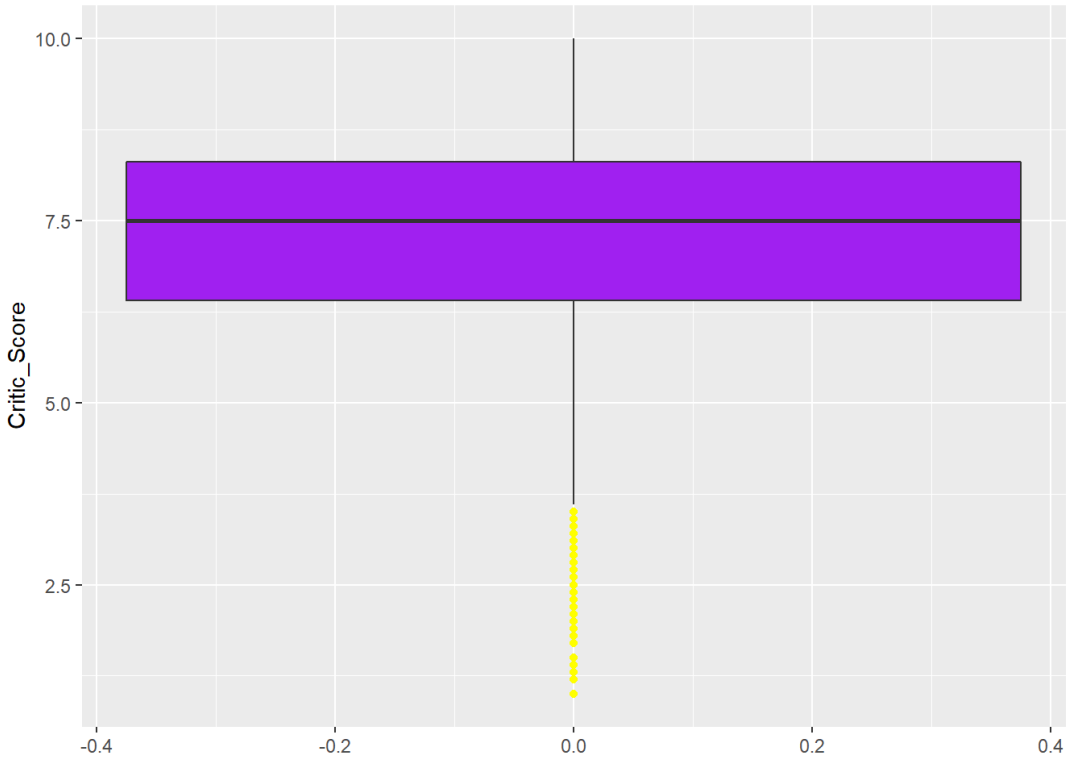

PC
PS2
DS
PS
XBL
GB
PS4
SNES
NES
PSP
GBA
3DS
PS3
NS
XOne
PSN
And
OSX
Wii
SAT
PSV
GEN
X360

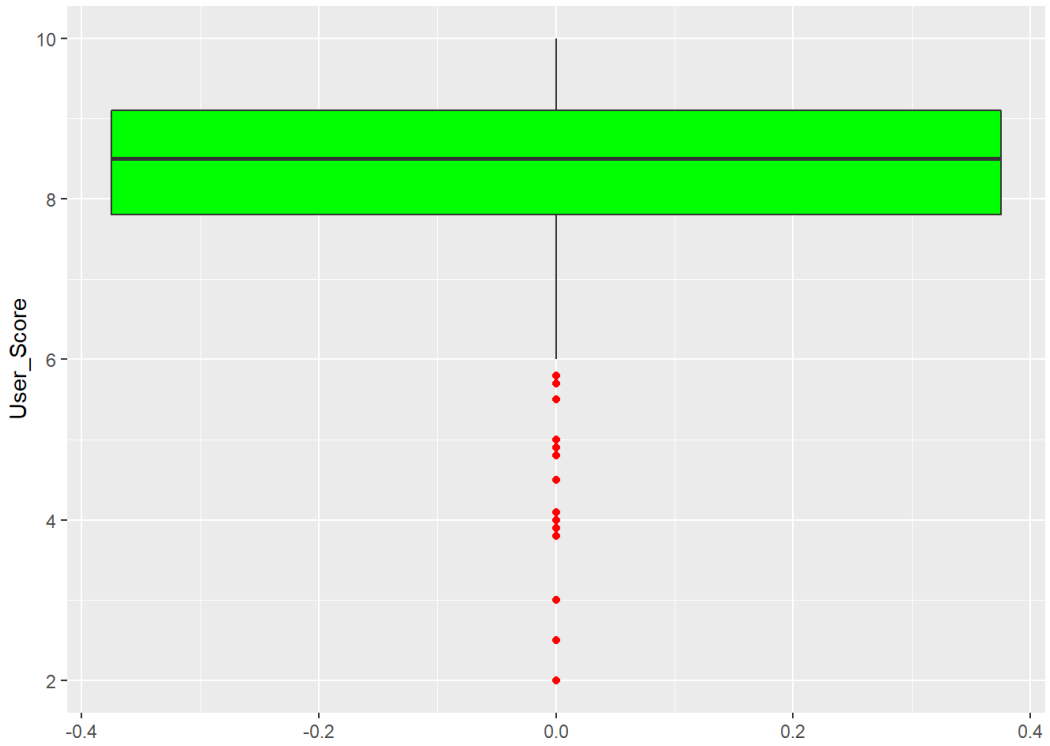

Misc
Action
Shooter
Simulation
Strategy
Fighting
Misc
Adventure
Sports
Racing
Role-Playing
Shooter
Platform
Simulation
Puzzle
Misc
Racing
Fighting
Role-Playing
Visual Novel
Strategy
Action
Role-Playing
Shooter
Action-Adventure
Racing
Platform
Shooter
Action
Strategy
Sports
Simulation
Adventure
Puzzle
Adventure
Action
Action
Racing
Platform
Strategy
Adventure
Adventure
Role-Playing
Sports
Misc
Sports
Role-Playing

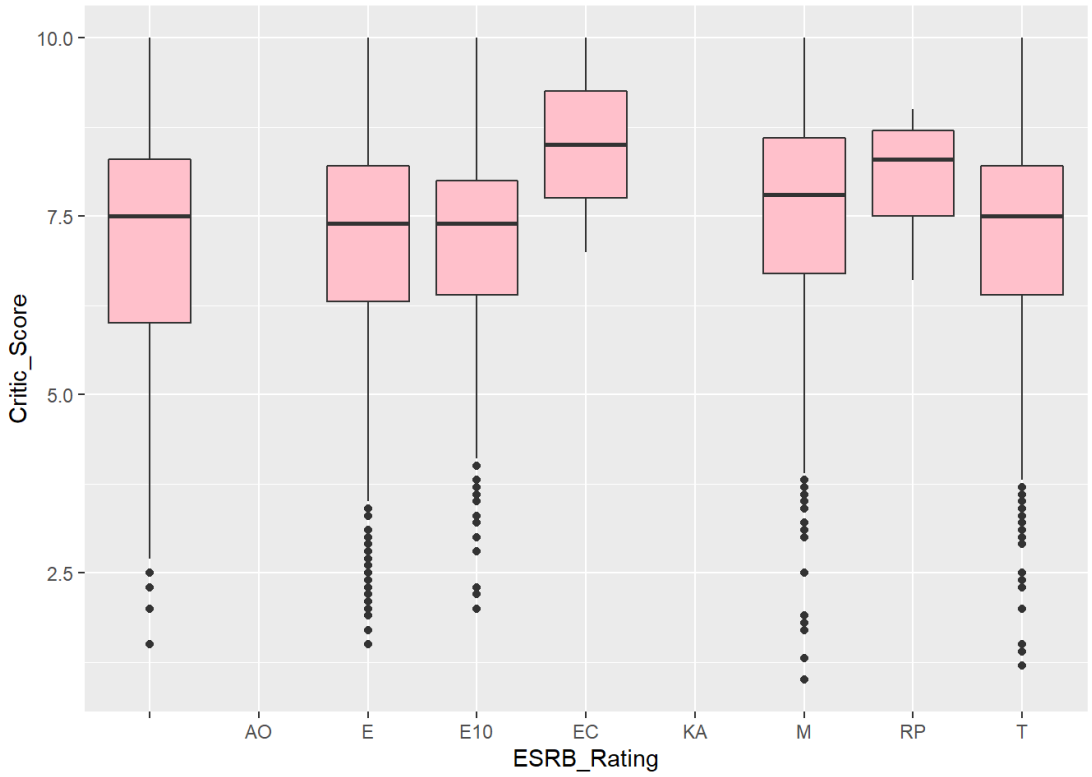

4000
3000
2000
1000
0
count
1970
1980
1990
2000
2010
2020
Year

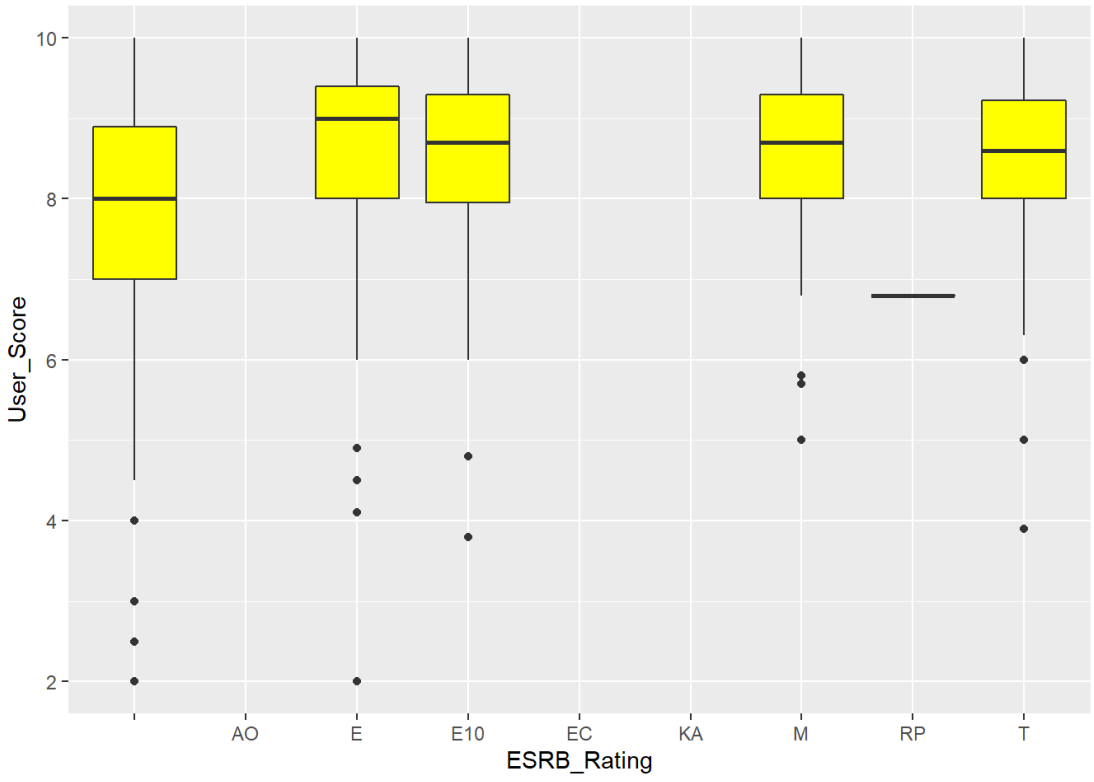

count
400
200
0
0
20
40
60
80
Total_Shipped

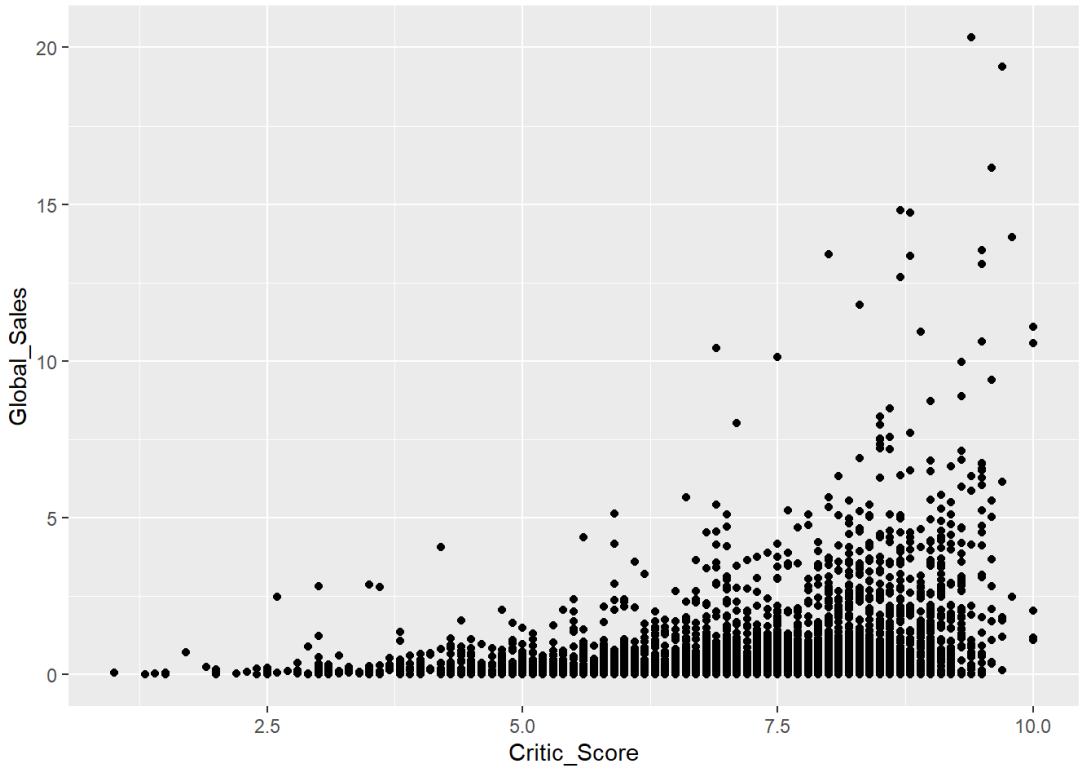

count
6000
4000
2000
0
0
5
10
15
20
Global_Sales

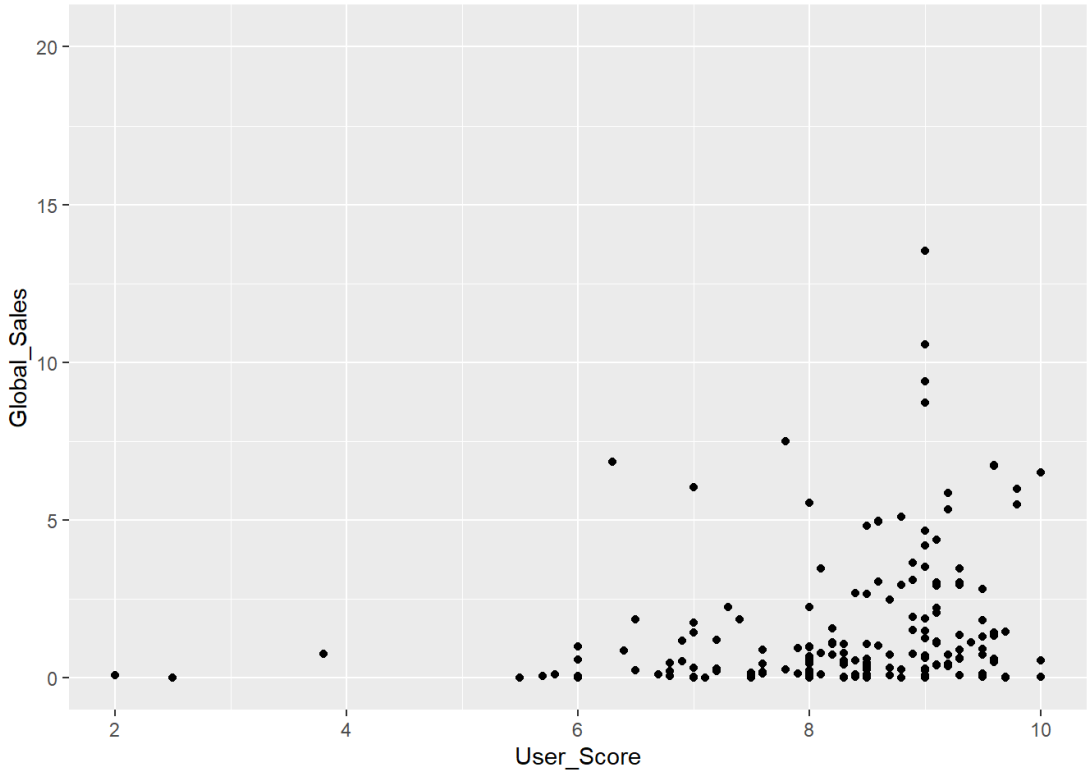

Critic_Score
10.0
7.5
5.0
2.5
-0.4
-0.2
0.0
0.2
0.4

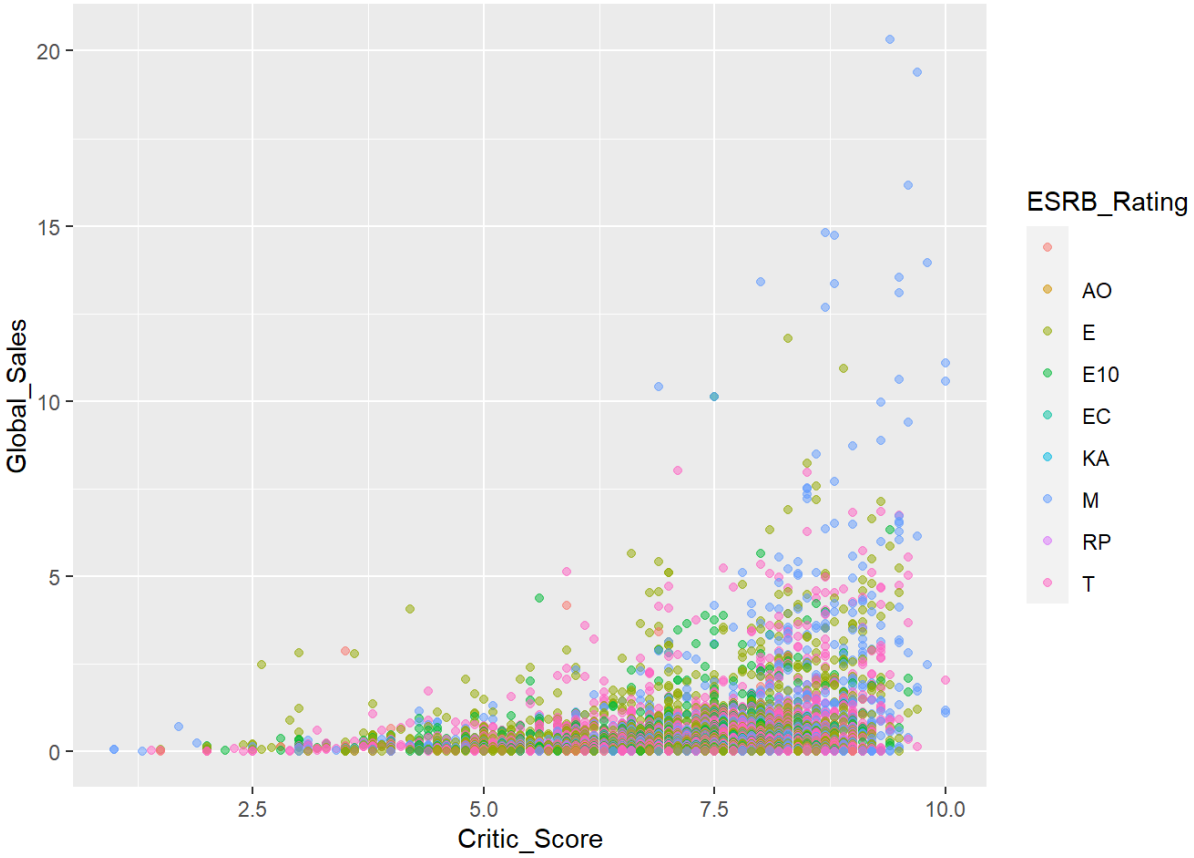

User_Score
10
8
6
4
2
-0.4
-0.2
0.0
0.2
0.4

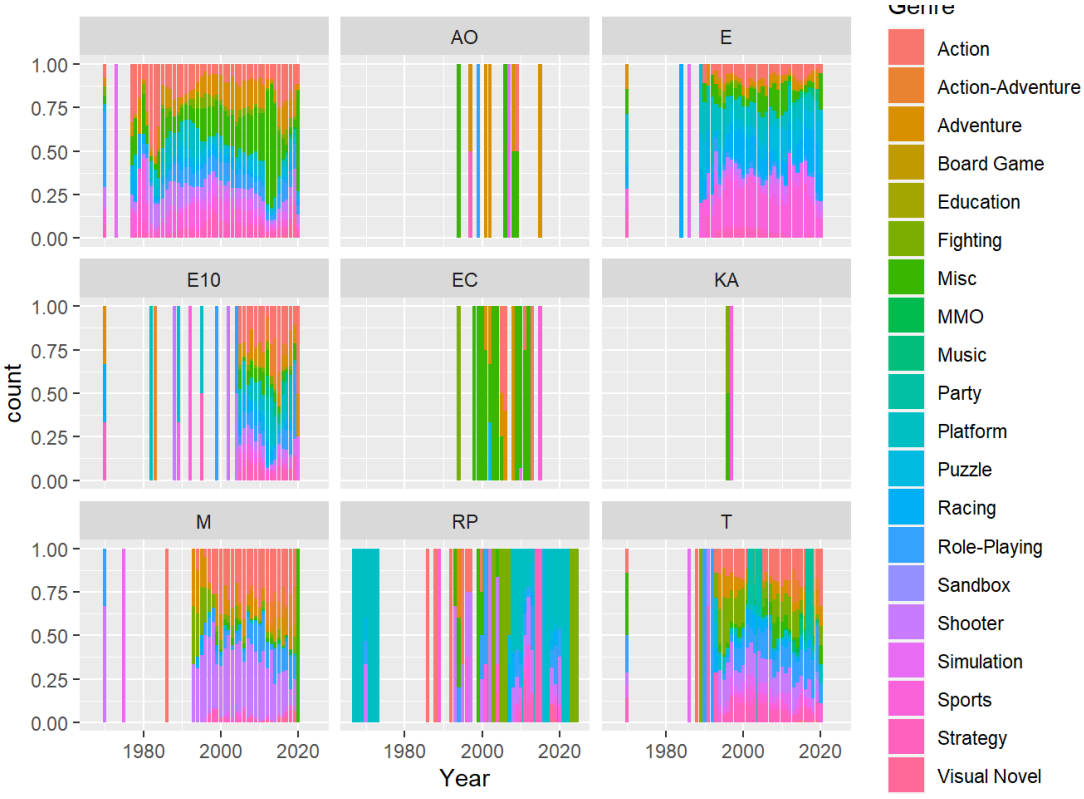

10.0
7.5
5.0
2.5
Critic_Score
AO
E
E10
EC
KA
M
RP
T
ESRB_Rating


10
8
6
4
2
User_Score
AO
E
E10
EC
KA
M
RP
T
ESRB_Rating


Global_Sales
20
15
10
5
0
2.5
5.0
7.5
10.0
Critic_Score


Global_Sales
20
15
10
5
0
2
4
6
8
10
User_Score


Global_Sales
Critic_Score
ESRB_Rating
AO
E
E10
EC
KA
M
RP
T


AO
E
E10
EC
KA
M
RP
T
count
Year
Genre
Action
Action-Adventure
Adventure
Board Game
Education
Fighting
Misc
MMO
Music
Party
Platform
Puzzle
Racing
Role-Playing
Sandbox
Shooter
Simulation
Sports
Strategy
Visual Novel

```
summary(dataSet$Critic_Score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00    6.40    7.50    7.21    8.30   10.00   49256
```

```
summary(dataSet$User_Score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    2.00    7.80    8.50    8.25    9.10   10.00   55457
```

```
summary(dataSet$Total_Shipped)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.03    0.20    0.59    1.89    1.80   82.86   53965
```

```
summary(dataSet$Global_Sales)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    0.03    0.12    0.37    0.36   20.32   36377
```

```
table(dataSet$Genre, dataSet$ESRB_Rating)
```

```
##
##                      AO    E  E10   EC   KA    M   RP    T
##   Action           4120    1  844  617    7    0  924   73 1081
##   Action-Adventure  241    0   15  119    0    0  139   22   73
##   Adventure        3620   10  567  272    6    0  337   25  456
##   Board Game         12    0    1    0    0    0    0    3    0
##   Education          10    0    2    0    0    0    0    0    0
##   Fighting         1041    0   90   53    1    1  114   11  774
##   Misc             7587    5 1175  210   37    1   66   23  372
##   MMO                40    0    0    0    0    0   10    1   23
##   Music              65    0   27   51    0    0    0    1   51
##   Party              42    0   13   15    0    0    0    1    4
##   Platform         1725    0 1177  359    0    0   12    8  164
##   Puzzle           1776    0 1230  102    1    0    5   13   35
##   Racing           1053    0 1437  194    0    0   33   17  296
##   Role-Playing     2711    1  261  251    0    0  432   52  843
##   Sandbox             5    0    0    1    0    0    0    0    3
##   Shooter          2125    0  335  160    0    0 1105   58  803
##   Simulation       1461    1  773  118    1    0   23    8  352
##   Sports           2310    0 2450  153    1    1   22   27  280
##   Strategy         1997    1  414  221    0    0   74   21  538
##   Visual Novel      228    0    0    1    0    0   18    4    9
```

```
table(dataSet$Genre, dataSet$Platform)
```

```
##
##                  2600  3DO  3DS 5200 7800  Aco ACPC   AJ Amig  And ApII  Arc
##   Action          302   12  192   26   21    0    6    4    4  133    0    0
##   Action-Adventure  0    0   33    0    0    0    0    0    0    8    0    0
##   Adventure         3   43   58    2    0    0    1    3    1   20    0    0
##   Board Game        0    0    0    0    0    0    0    0    0    0    0    0
##   Education         0    0    1    0    0    0    0    0    0    0    0    0
##   Fighting          3   13   17    0    3    0    1    4    1    4    0   10
##   Misc             21   87  423    0    0    1    3    3   16  683    4   29
##   MMO               0    0    1    0    0    0    0    0    0    0    0    0
##   Music             0    0   10    0    0    0    0    0    0    1    0    0
##   Party             0    0    6    0    0    0    0    0    0    0    0    0
##   Platform         17    7   71    9    6    0    4    9    2    8    0    0
##   Puzzle           19   13   74    6    3    0    0    5    1   32    0    0
##   Racing           19   10   19    2    3    0    0    9    0   10    0    0
##   Role-Playing      4   10  123    0    0    0    2    1    0   38    0    0
##   Sandbox           0    0    0    0    0    0    0    0    0    0    0    0
##   Shooter          68   37   19   20    9    0    1   14    4   16    0    0
##   Simulation        4   23   71    1    2    0    1    8    0   22    0    0
##   Sports           39   35   49    7   12    0    0    7    0   13    0    0
##   Strategy          2   14   24    0    0    0    1    5    0   19    0    0
##   Visual Novel      0    0    1    0    0    0    0    0    0    0    0    0
##
```

```
table(dataSet$Genre, dataSet$Year)
```

```
##
##                  1970 1973 1975 1977 1978 1979 1980 1981 1982 1983 1984 1985
##   Action           12    0    0    4   14    1    3   10  103  158   31   11
##   Action-Adventure  0    0    0    0    0    0    0    0    4    2    0    0
##   Adventure         6    0    0    1    1    1    1    3    6   13   12   11
##   Board Game        0    0    0    0    0    0    0    0    0    0    0    0
##   Education         0    0    0    0    0    0    0    0    0    0    0    0
##   Fighting          3    0    0    0    0    0    2    0    0    0    1    2
##   Misc              9    0    0    2   10    0    8    4    6    9   13   21
##   MMO               0    0    0    0    0    0    0    0    0    0    0    0
##   Music             0    0    0    0    0    0    0    0    0    1    0    0
##   Party             0    0    0    0    0    0    0    0    0    0    0    0
##   Platform          2    0    0    0    1    0    0    0   12   29    6    8
##   Puzzle            6    0    0    0    6    0    0    1   12   17    3    7
##   Racing            2    0    0    2    6    0    2    1    2    9    3    4
##   Role-Playing     32    0    0    0    0    1    2    1    6    3    2    8
##   Sandbox           0    0    0    0    0    0    0    0    0    0    0    0
##   Shooter          15    0    0    1    3    0    1    9   44   44   11   10
##   Simulation        2    1    1    0    0    0    0    1    8    0    0    3
##   Sports            6    0    0    2    6    1   13    6    6   12    3    7
##   Strategy          9    0    0    0    1    1    3    1    5    2    3    1
##   Visual Novel      0    0    0    0    0    0    0    0    0    1    0    0
##
```

```
table(dataSet$ESRB_Rating, dataSet$Platform)
```

```
##
##       2600  3DO  3DS 5200 7800  Aco ACPC   AJ Amig  And ApII  Arc  AST BBCM
##        501  256  781   73   59    1   20   26   29  986    4   37   17    1
##   AO     0    1    0    0    0    0    0    0    0    0    0    0    0    0
##   E      0   26  183    0    0    0    0   38    0    7    0    0    0    0
##   E10    0    0  121    0    0    0    0    0    0    6    0    0    0    0
##   EC     0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   KA     0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   M      0    6   21    0    0    0    0    3    0    2    0    0    0    0
##   RP     0    0    8    0    0    0    0    0    0    0    0    0    0    0
##   T      0   15   78    0    0    0    0    5    0    6    0    2    0    0
##
##       BRW C128  C64 CD32  CDi   DC   DS  DSi DSiW  FMT   GB  GBA  GBC   GC
##        63    1   31    3    5  403 1538   76  343    3 1107  662    5   99
##   AO    0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   E     0    0    0    0    0  129 1228    0  332    0  469  830   12  288
##   E10   0    0    0    0    0    0  318    0   48    0    0   53    0   35
##   EC    0    0    0    0    0    0    9    0    1    0    4    4    0    0
##   KA    0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   M     0    0    0    0    1   34   12    0    0    0    2    9    0   44
##   RP    0    0    0    0    0    2   20    0    3    0    0    0    0    0
##   T     0    0    0    0    0   87  167    0   26    0   19  100    0  200
##
##       GEN   GG  GIZ  Int  iOS iQue Linux Lynx  Mob   MS  MSD  MSX  N64  NES
##       684  271   43  134   47   17  310   87   21  336   27    7   93 1102
##   AO    0    0    0    0    0    0    1    0    0    0    0    0    0    0
##   E   105   59   10    0    4    0    2    2    0    0    0    0  215    2
##   E10   0    0    0    0    2    0   16    0    0    0    0    0    0    0
##   EC    0    0    0    0    0    0    0    0    0    0    0    0    2    1
##   KA    0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   M     3    0    0    0    2    0    6    0    0    0    0    0   28    0
##   RP    0    0    0    0    1    0    2    0    0    0    0    0    0    0
##   T    13    4    2    0    1    0   13    0    1    0    0    0   55    0
##
```

```
table(dataSet$Platform, dataSet$Year)
```

```
##
##         1970 1973 1975 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987
##   2600     1    0    0   10   45    2   10   20  145  179   24    1    2    9
##   3DO      0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   3DS      0    0    0    0    0    0    0    0    0    1    0    0    0    0
##   5200     0    0    0    0    0    0    0    1    8   38   22    0    0    2
##   7800     0    0    0    0    0    0    0    0    0    0    0    0    3   16
##   Aco      0    0    0    0    0    0    0    0    0    0    1    0    0    0
##   ACPC     0    0    0    0    0    0    0    0    0    0    1    4    2    7
##   AJ       0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   Amig     0    0    0    0    0    0    0    0    0    0    0    0    2    3
##   And      0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   ApII     0    0    0    0    0    0    0    0    0    0    1    2    0    0
##   Arc      0    0    0    0    0    0    1    1    1    0    0    1    0    2
##   AST      0    0    0    0    0    0    0    0    0    0    0    2    0    2
##   BBCM     0    0    0    0    0    0    0    0    0    0    1    0    0    0
##   BRW      0    0    0    0    0    0    0    0    0    0    0    0    0    1
##   C128     0    0    0    0    0    0    0    0    0    0    0    1    0    0
##   C64      0    0    0    0    0    0    0    0    0    1    7    5    4    5
##   CD32     0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   CDi      0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   DC       0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   DS       3    0    1    0    0    0    0    0    0    1    0    1    1    0
##   DSi      0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   DSiW     0    0    0    1    0    0    0    1    0    0    0    0    0    0
##   FMT      0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   GB       0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   GBA      1    0    0    0    0    0    0    0    0    0    0    0    0    0
##   GBC      0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   GC       0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   GEN      0    0    0    0    0    0    0    0    0    0    0    0    2    1
##   GG       0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   GIZ      0    0    0    0    1    0    0    0    0    0    0    0    0    0
##   Int      0    0    0    0    0    0   19    7   29   47    0    2    5   10
```

```
table(dataSet$ESRB_Rating, dataSet$Year)
```

```
##
##       1970 1973 1975 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987
##         61    1    0   12   48    5   35   37  213  299   87   93  141  259
##   AO     0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   E      7    0    0    0    0    0    0    0    0    0    1    0    1    0
##   E10    3    0    0    0    0    0    0    0    1    1    0    0    0    0
##   EC     0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   KA     0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   M      3    0    1    0    0    0    0    0    0    0    0    0    1    0
##   RP    16    0    0    0    0    0    0    0    0    0    0    0    2    0
##   T     14    0    0    0    0    0    0    0    0    0    0    0    1    0
##
##       1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001
##        279  426  656  775  916 1001 1023  710  681  602  677  704  768  729
##   AO     0    0    0    0    0    0    1    0    0    2    0    1    0    3
##   E      0    5    9    8   16   26   97  347  375  267  260  383  514  487
##   E10    1    3    0    0    1    0    0    2    0    0    0    1    0    0
##   EC     0    0    0    0    0    0    1    0    0    0    2    4    1    4
##   KA     0    0    0    0    0    0    0    0    2    1    0    0    0    0
##   M      0    0    0    0    0    3   16   18   40   55   38   54   74   61
##   RP     1    1    0    0    1    6    5    3    8    4    0    1    4    3
##   T      1    1    1    3    1    7   20   82  114   99  117  137  200  206
##
##       2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
##        708  684  681  659  855 1031 1176 2172 1496 1714 1084 1294 2340 1105
##   AO     2    0    0    0    2    1    2    2    0    0    0    0    0    3
##   E    593  540  440  467  555  777  917 1080  971  683  131  118  126  127
##   E10    1    0    2  136  208  259  301  415  344  319   83  101  145  106
##   EC     3    1    1    4    5    0    1    5   14    4    1    1    0    1
##   KA     0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   M     98  120  147  208  141  139  162  295  295  302  129  123  160  136
##   RP     1    0    6    0    2    6    5   19   15   17   18   12    2    1
##   T    330  393  334  347  344  360  415  519  526  450  114   74  132  154
##
```

```
dataSet%>%
filter(!is.na(dataSet$Global_Sales))%>%
group_by(Platform)%>%
summarize(averageGlobalSales=mean(Global_Sales, na.rm = T))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 39 x 2
##    Platform averageGlobalSales
##    <chr>                 <dbl>
##  1 2600                  0.676
##  2 3DO                   0.0475
##  3 3DS                   0.199
##  4 DC                    0.273
##  5 DS                    0.193
##  6 GB                    0.489
##  7 GBA                   0.255
##  8 GBC                   1.45
##  9 GC                    0.229
## 10 GEN                   0.697
## # ... with 29 more rows
```

```
dataSet%>%
filter(!is.na(dataSet$Global_Sales))%>%
group_by(ESRB_Rating)%>%
summarize(averageGlobalSales=mean(Global_Sales, na.rm = T))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 7 x 2
##   ESRB_Rating averageGlobalSales
##   <chr>                   <dbl>
## 1 ""                      0.144
## 2 "E"                     0.397
## 3 "E10"                   0.388
## 4 "EC"                    0.198
## 5 "M"                     0.783
## 6 "RP"                    0.154
## 7 "T"                     0.409
```

```
dataSet%>%
filter(!is.na(dataSet$Global_Sales))%>%
group_by(Year)%>%
summarize(averageGlobalSales=mean(Global_Sales, na.rm = T))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 46 x 2
##     Year averageGlobalSales
##    <dbl>              <dbl>
##  1  1970              0.116
##  2  1977              0.833
##  3  1978              0.688
##  4  1979              0.31
##  5  1980              0.43
##  6  1981              1.50
##  7  1982              0.776
##  8  1983              0.613
##  9  1984              0.485
## 10  1985              0.548
## # ... with 36 more rows
```

```
dataSet%>%
filter(!is.na(dataSet$Critic_Score))%>%
group_by(Platform)%>%
summarize(averageCriticScore = mean(Critic_Score, na.rm = T))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 36 x 2
##    Platform averageCriticScore
##    <chr>                 <dbl>
##  1 3DS                    7.07
##  2 AJ                     4
##  3 DC                     7.06
##  4 DS                     7.13
##  5 DSiW                   6.39
##  6 GB                     7.64
##  7 GBA                    6.59
##  8 GBC                    9.1
##  9 GC                     6.99
## 10 GEN                    8.67
## # ... with 26 more rows
```