Videogames Analysis With R by Damien

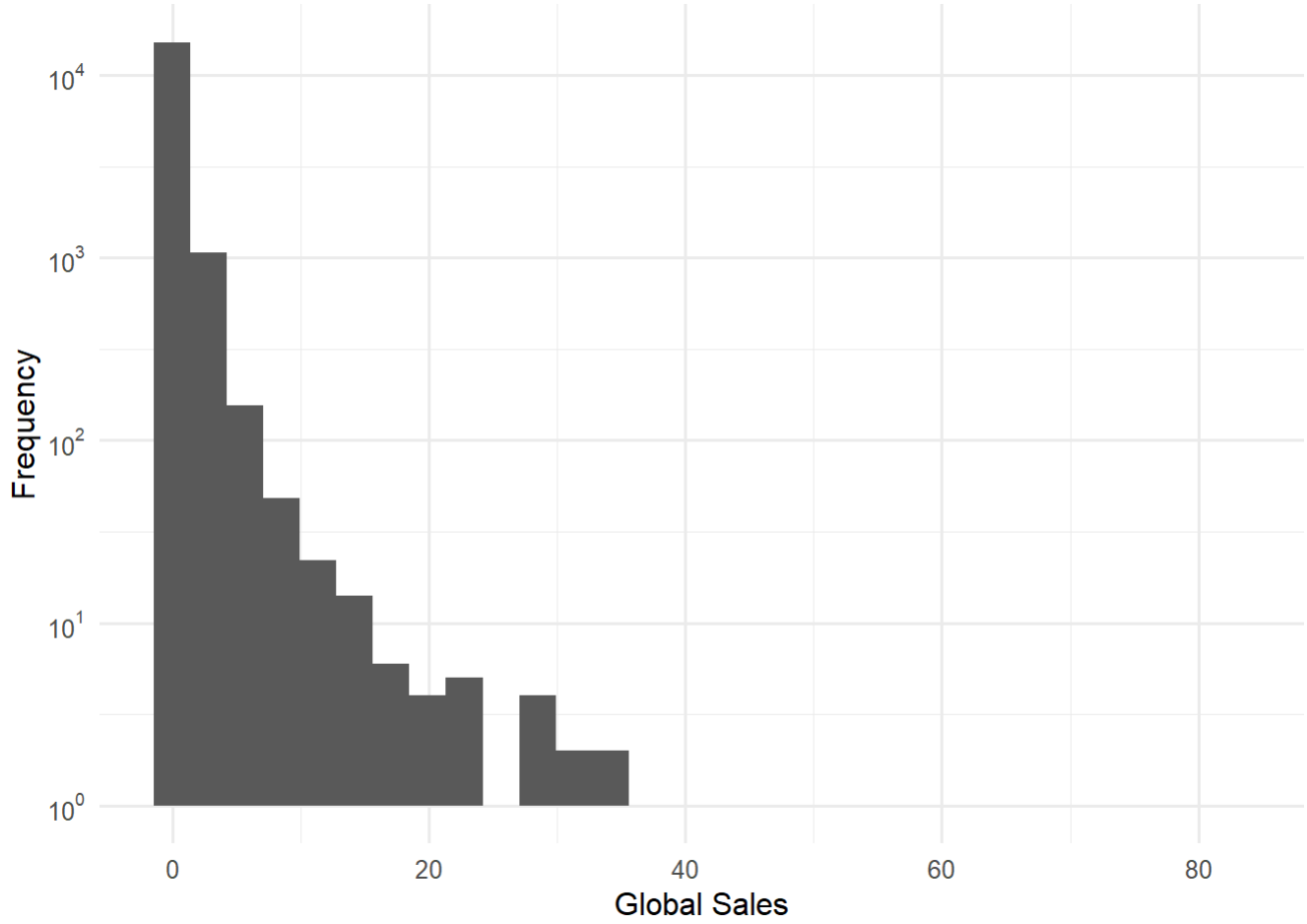
This report explores a dataset containing regional and global sales, along with user and critic scores for approximately 16,700 video games.

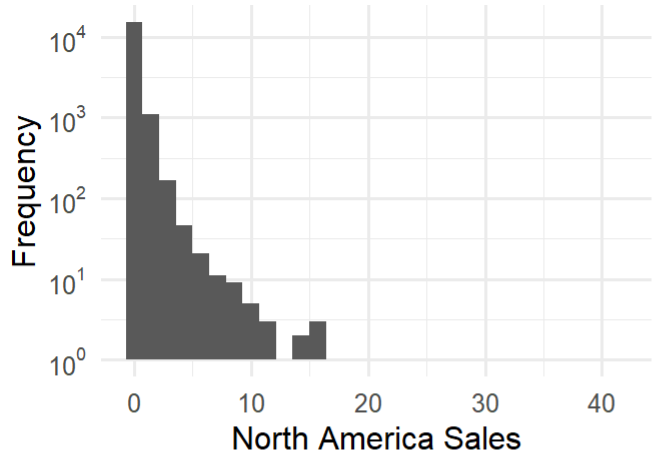
Univariate Plots Section

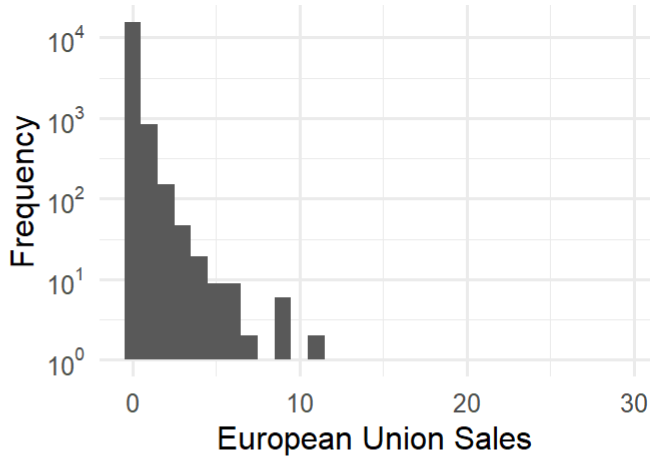
```
## 'data.frame':    16719 obs. of  16 variables:
##  $ Name           : Factor w/ 11563 levels "","'98 Koshien",..: 11059 9406 5573 11061 7417 97
71 6693 11057 6696 2620 ...
##  $ Platform       : Factor w/ 31 levels "2600","3DO","3DS",..: 26 12 26 26 6 6 5 26 26 12 ...
##  $ Year_of_Release: Factor w/ 40 levels "1980","1981",..: 27 6 29 30 17 10 27 27 30 5 ...
##  $ Genre          : Factor w/ 13 levels "","Action","Adventure",..: 12 6 8 12 9 7 6 5 6 10
...
##  $ Publisher      : Factor w/ 582 levels "10TACLE Studios",..: 371 371 371 371 371 371 371 37
1 371 371 ...
##  $ NA_Sales       : num  41.4 29.1 15.7 15.6 11.3 ...
##  $ EU_Sales       : num  28.96 3.58 12.76 10.93 8.89 ...
##  $ JP_Sales       : num  3.77 6.81 3.79 3.28 10.22 ...
##  $ Other_Sales    : num  8.45 0.77 3.29 2.95 1 0.58 2.88 2.84 2.24 0.47 ...
##  $ Global_Sales   : num  82.5 40.2 35.5 32.8 31.4 ...
##  $ Critic_Score   : int  76 NA 82 80 NA NA 89 58 87 NA ...
##  $ Critic_Count   : int  51 NA 73 73 NA NA 65 41 80 NA ...
##  $ User_Score     : Factor w/ 97 levels "","0","0.2","0.3",..: 79 1 82 79 1 1 84 65 83 1 ...
##  $ User_Count     : int  322 NA 709 192 NA NA 431 129 594 NA ...
##  $ Developer      : Factor w/ 1697 levels "","10tacle Studios",..: 1035 1 1035 1035 1 1 1035
1035 1035 1 ...
##  $ Rating         : Factor w/ 9 levels "","AO","E","E10+",..: 3 1 3 3 1 1 3 3 3 1 ...
```

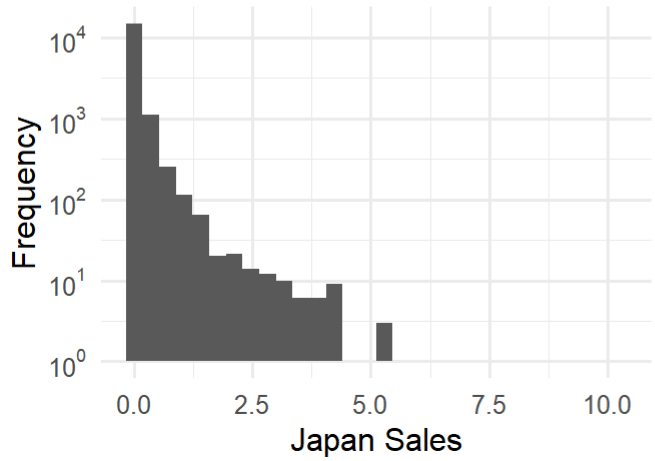
```
##                           Name          Platform    Year_of_Release
##  Need for Speed: Most Wanted:   12   PS2    :2161   2008   :1427
##  FIFA 14                     :    9   DS     :2152   2009   :1426
##  LEGO Marvel Super Heroes    :    9   PS3    :1331   2010   :1255
##  Madden NFL 07               :    9   Wii    :1320   2007   :1197
##  Ratatouille                 :    9   X360   :1262   2011   :1136
##  Angry Birds Star Wars       :    8   PSP    :1209   2006   :1006
##  (Other)                     :16663   (Other):7284   (Other):9272
##           Genre                              Publisher
##  Action      :3370   Electronic Arts             : 1356
##  Sports      :2348   Activision                  :  985
##  Misc        :1750   Namco Bandai Games          :  939
##  Role-Playing:1500   Ubisoft                     :  933
##  Shooter     :1323   Konami Digital Entertainment:  834
##  Adventure   :1303   THQ                         :  715
##  (Other)     :5125   (Other)                     :10957
##     NA_Sales          EU_Sales         JP_Sales        Other_Sales
##  Min.   : 0.0000   Min.   : 0.000   Min.   : 0.0000   Min.   : 0.00000
##  1st Qu.: 0.0000   1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.: 0.00000
##  Median : 0.0800   Median : 0.020   Median : 0.0000   Median : 0.01000
##  Mean   : 0.2633   Mean   : 0.145   Mean   : 0.0776   Mean   : 0.04733
##  3rd Qu.: 0.2400   3rd Qu.: 0.110   3rd Qu.: 0.0400   3rd Qu.: 0.03000
##  Max.   :41.3600   Max.   :28.960   Max.   :10.2200   Max.   :10.57000
##
##   Global_Sales      Critic_Score    Critic_Count      User_Score
##  Min.   : 0.0100   Min.   :13.00   Min.   :  3.00          :6704
##  1st Qu.: 0.0600   1st Qu.:60.00   1st Qu.: 12.00   tbd    :2425
##  Median : 0.1700   Median :71.00   Median : 21.00   7.8    : 324
##  Mean   : 0.5335   Mean   :68.97   Mean   : 26.36   8      : 290
##  3rd Qu.: 0.4700   3rd Qu.:79.00   3rd Qu.: 36.00   8.2    : 282
##  Max.   :82.5300   Max.   :98.00   Max.   :113.00   8.3    : 254
##                    NA's   :8582    NA's   :8582     (Other):6440
##    User_Count          Developer        Rating
##  Min.   :    4.0            :6623          :6769
##  1st Qu.:   10.0   Ubisoft  : 204   E      :3991
##  Median :   24.0   EA Sports: 172   T      :2961
##  Mean   :  162.2   EA Canada: 167   M      :1563
##  3rd Qu.:   81.0   Konami   : 162   E10+   :1420
##  Max.   :10665.0   Capcom   : 139   EC     :   8
##  NA's   :9129      (Other)  :9252   (Other):   7
```

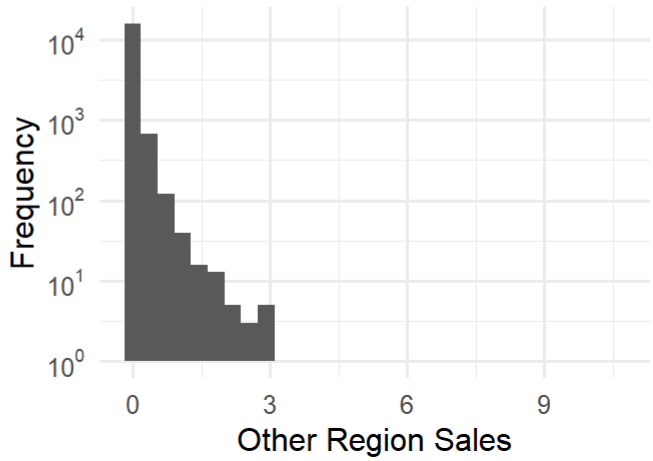
Our dataset contains 16 variables, with almost 17,000 observations.


Frequency
10^4
10^3
10^2
10^1
10^0
0
20
40
60
80
Global Sales


Frequency
10^4
10^3
10^2
10^1
10^0
0
10
20
30
40
North America Sales


Frequency
10^4
10^3
10^2
10^1
10^0
0
10
20
30
European Union Sales


Frequency
10^4
10^3
10^2
10^1
10^0
0.0
2.5
5.0
7.5
10.0
Japan Sales


Frequency
10^4
10^3
10^2
10^1
10^0
0
3
6
9
Other Region Sales

```
## [1] "Global Sales"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0100  0.0600  0.1700  0.5363  0.4700 82.5300
```

```
## [1] "North America Sales"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.080   0.264   0.240  41.360
```

```
## [1] "European Union Sales"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0200  0.1459  0.1100 28.9600
```

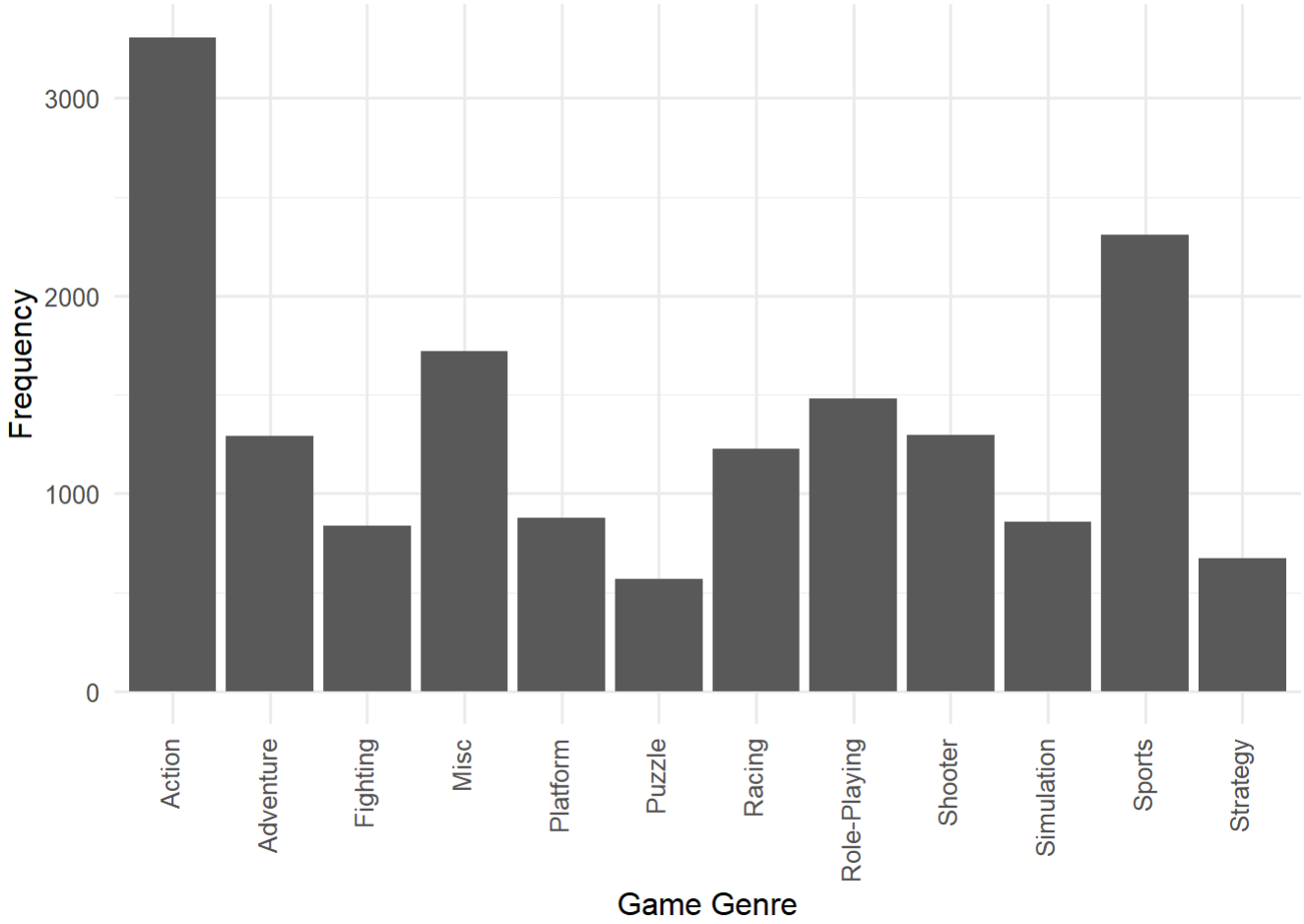
```
## [1] "Japan Sales"
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00000  0.00000  0.00000  0.07849  0.04000 10.22000
```

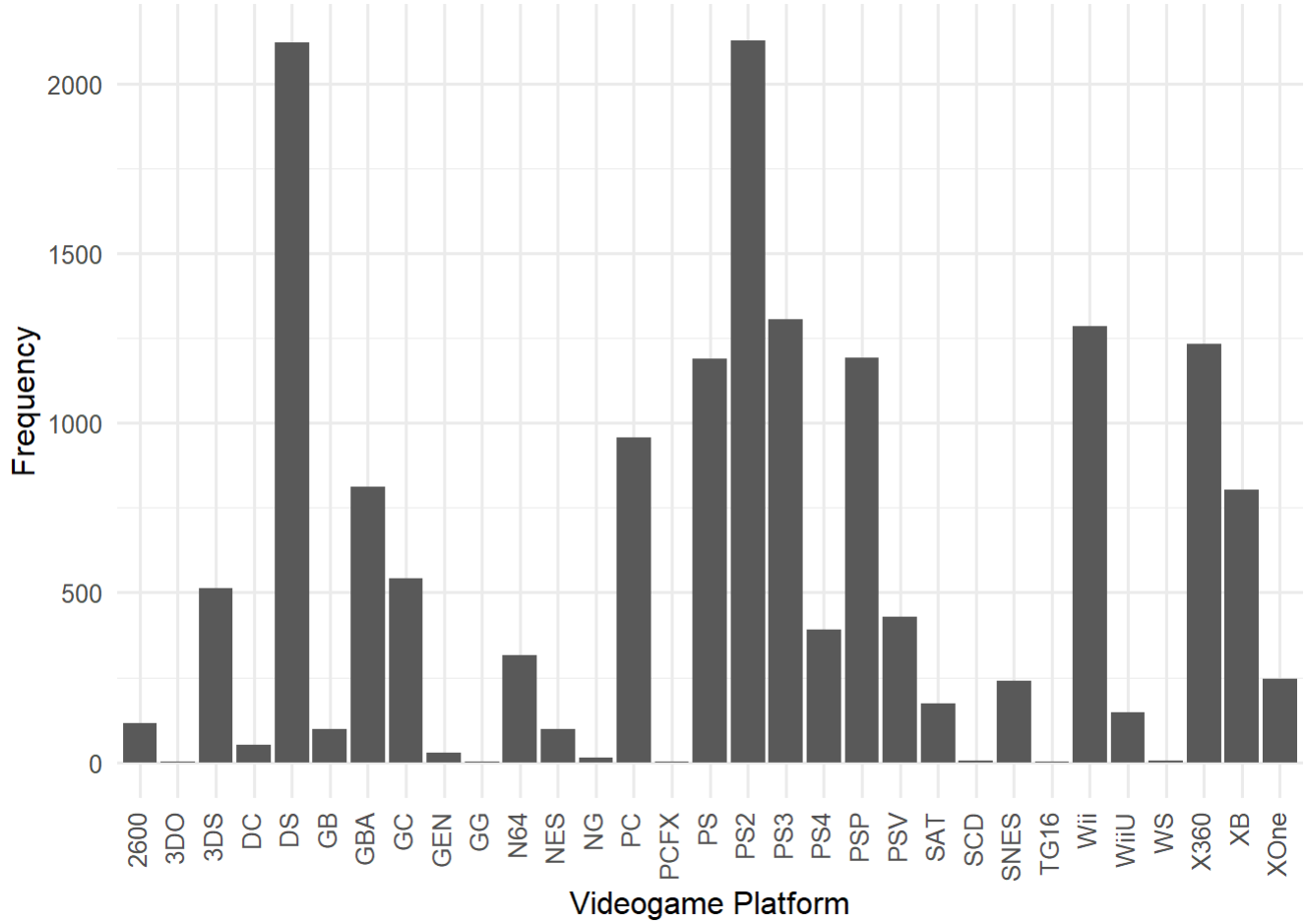
```
## [1] "Other Region Sales"
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00000  0.00000  0.01000  0.04759  0.03000 10.57000
```

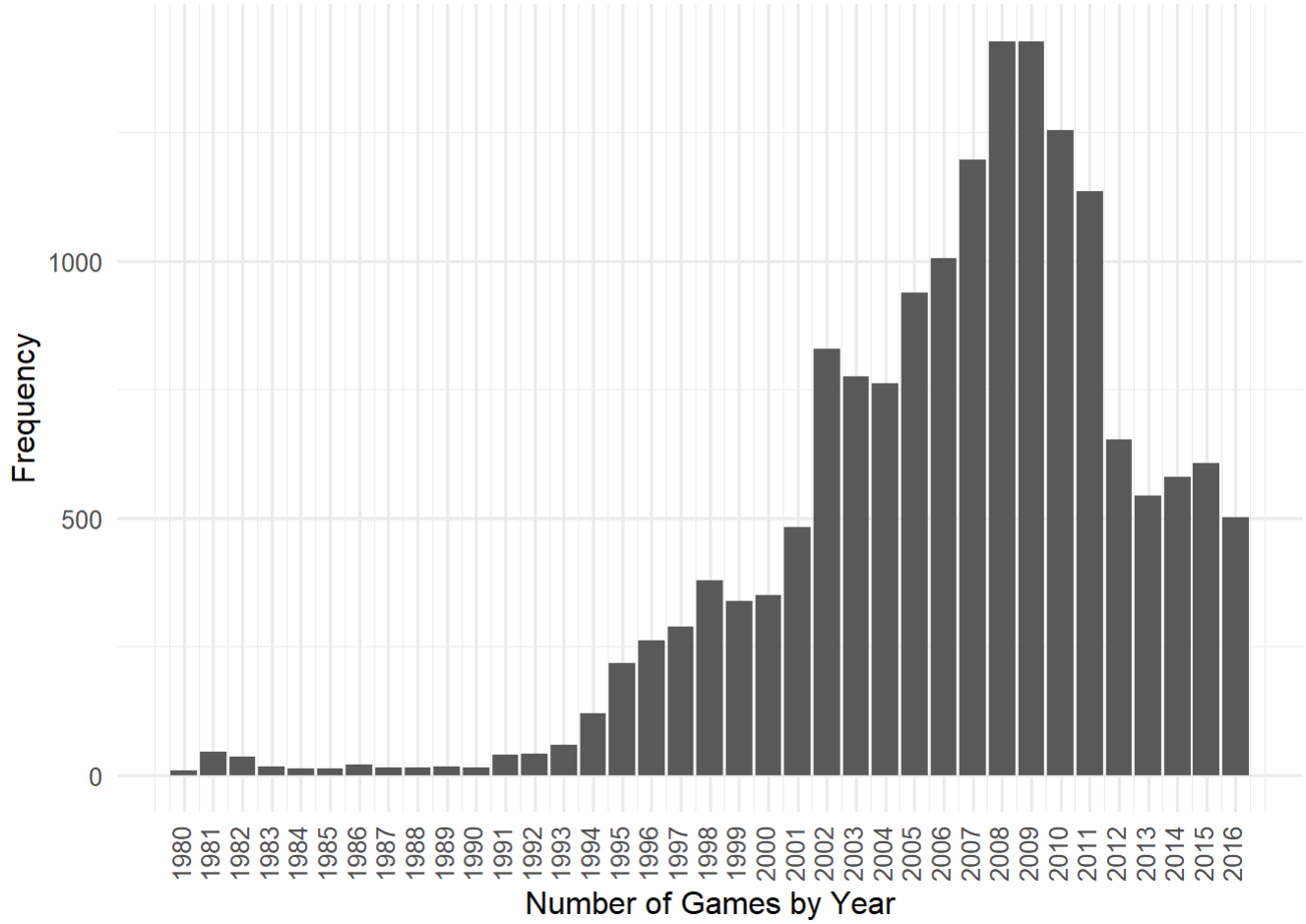
I wanted to view the distribution of sales among all the regions, as well as globally. All of the sales distributions are skewed to the right. Is their a specific genre that's causing North America to have more sales than the other regions? Why is their a gap in quantity of sales for most of the regions? I'm curious if this is due to outliers.



The Action Genre appears to be the most common, followed by the Sports genre. Genre has multiple peaks which makes it a multimodal distribution. Next I'll take a look at the distribution of the platforms.

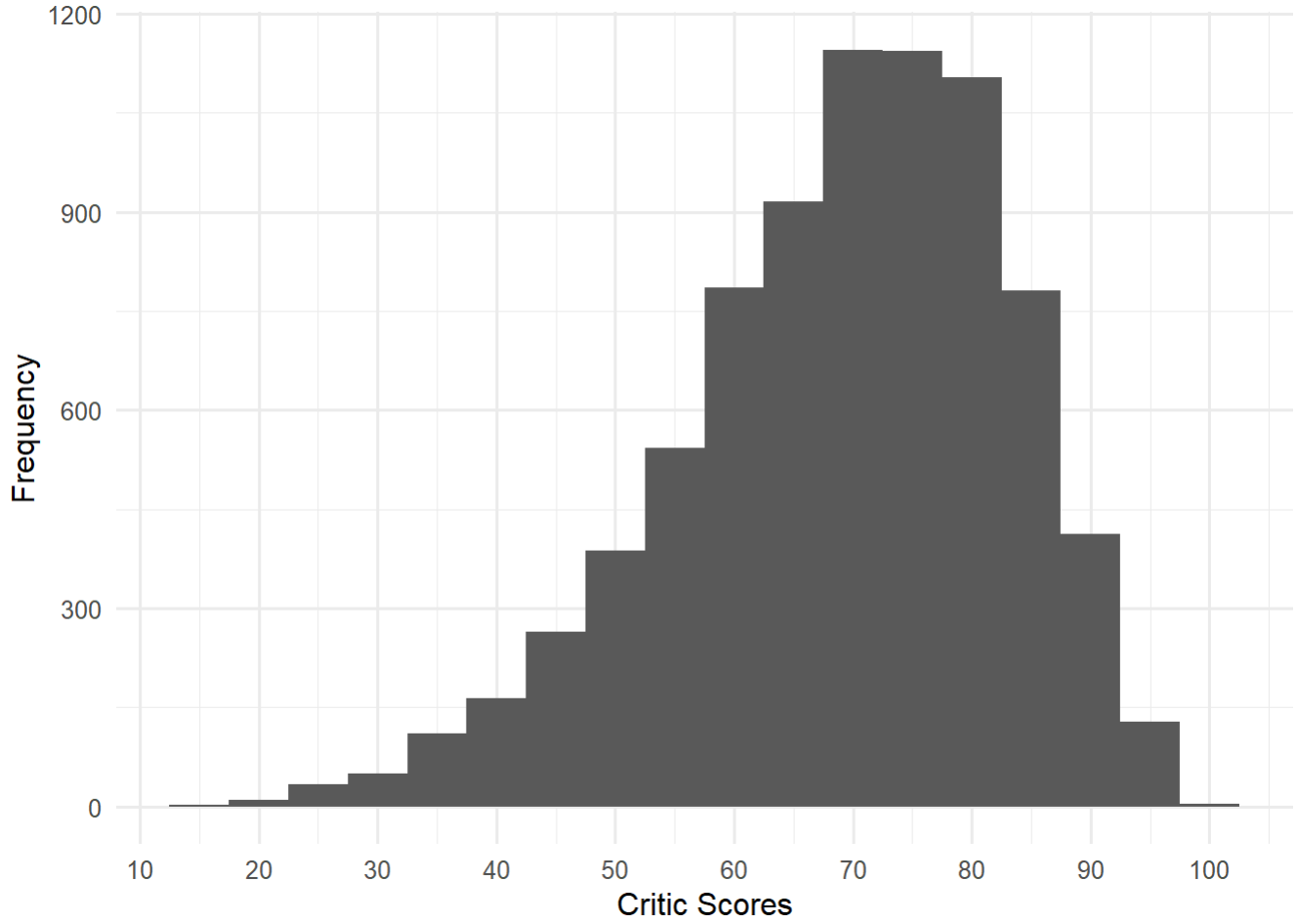


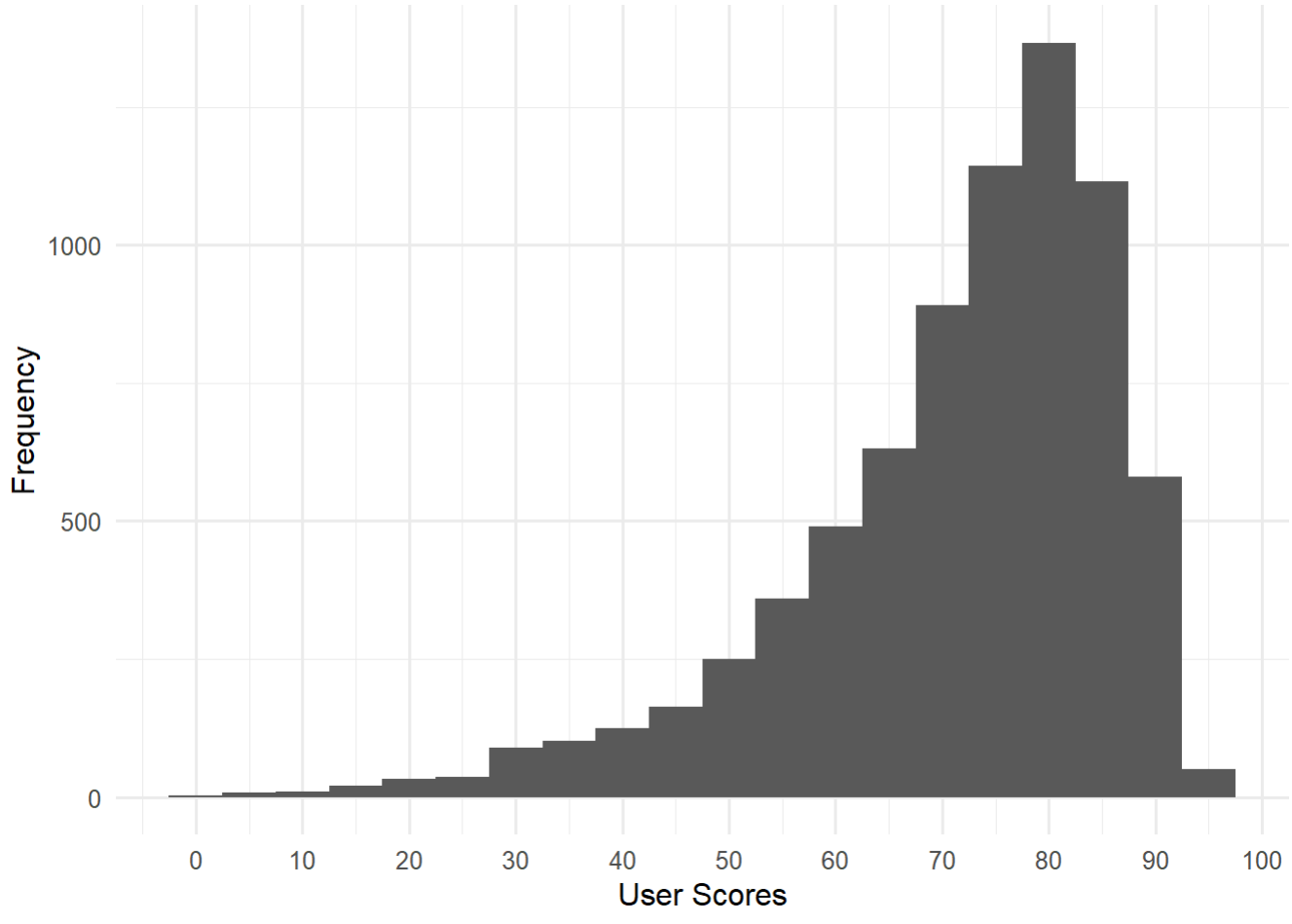
The platforms with the largest game library are the Nintendo DS and the Playstation 2. Platform also has multiple peaks making it a multimodal distribution. What does the distribution of games by year look?

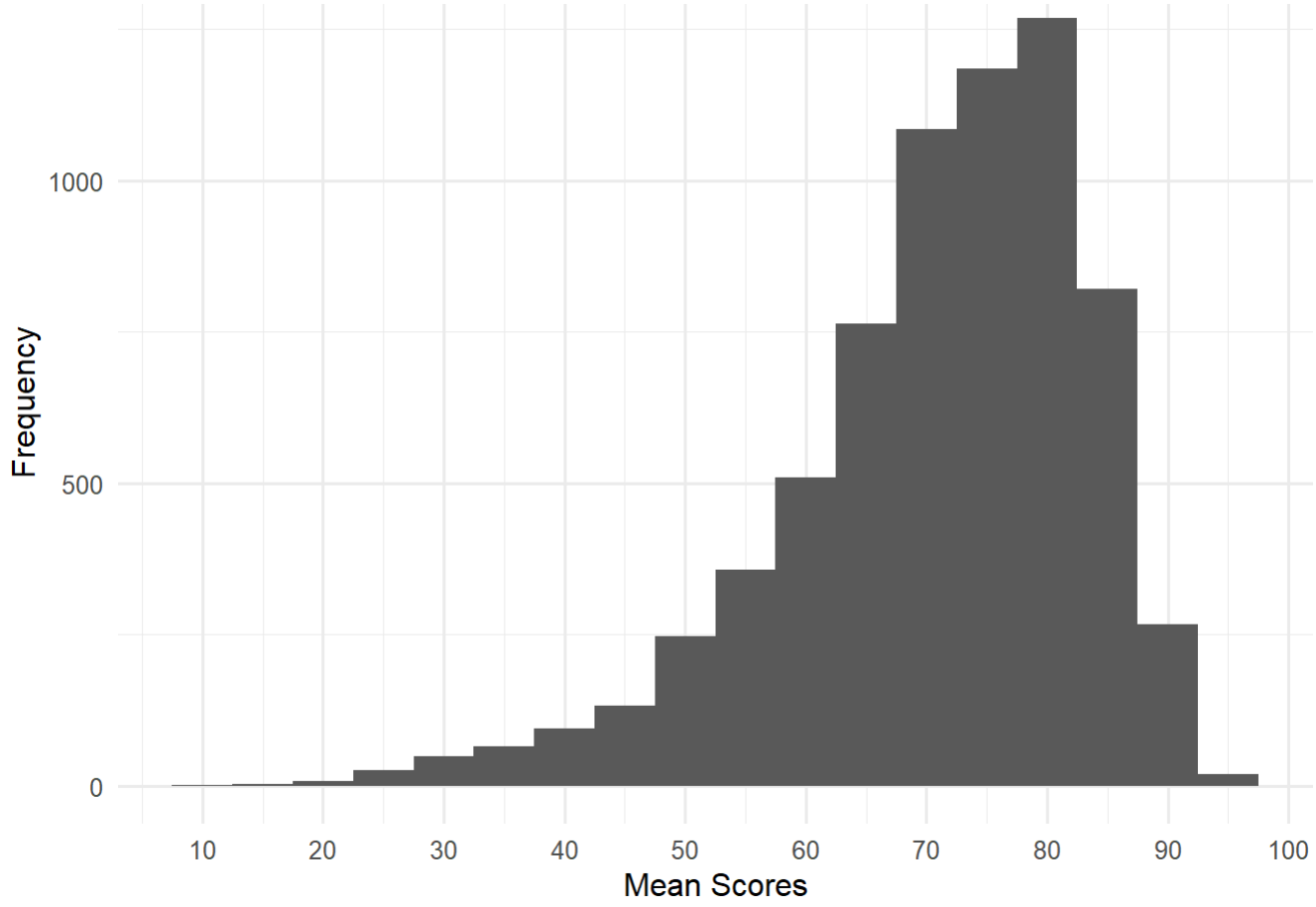


We can see a peak between the years 2008 and 2009. There were observations beyond the year 2016. The dataset was created in 2016 so there shouldn't be any observations beyond this year. I had to remove the data from 2017 and 2020 since there was no way to validate these observations at the time of creation.

Next, I want to take a look at the user and critic scores and see their distribution.


1200
900
600
300
0
Frequency
10
20
30
40
50
60
70
80
90
100
Critic Scores


1000
500
0
Frequency
0
10
20
30
40
50
60
70
80
90
100
User Scores




```
## [1] "Critic Score"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   13.00   60.00   71.00   68.99   79.00   98.00    8461
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    3.00   12.00   22.00   26.44   36.00  113.00    8461
```

```
## [1] "User Score"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   64.00   75.00   71.26   82.00   97.00    8981
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       4      10      24     163      81   10665    8981
```

```
## [1] "Mean Score"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   11.50   64.50   73.50   71.05   80.00   95.00    9550
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     8.0    30.0    55.0   203.2   128.0 10697.0    9550
```

On average, Critics rate games lower than Users. The Critics histogram is less left skewed than the users. There's a substantial number of user votes compared to the critic reviews, whic could cause the average of the two to inaccurate relationships. Does either one of them have a relationship with the different Sales variables?

Univariate Analysis

What is the structure of your dataset?

There are 16719 observations in the dataset with 16 features. The variables NA_Sales, EU_Sales, JP_Sales, Other_Sales, and Global_Sales are represented as millions of units. The Critic_Score and User_Score comes from the website Metacritic which was established in 2001.

Other observations:
* The year 2002 saw a sharp rise in video game releases
* The mean User Score is 71.26 while the mean Critic Score is 68.99
* The mean number of scores per game is 203.2
* North America has the largest amount of sales across the 4 regions
* The Action Genre seems to generate the most sales.

What is/are the main feature(s) of interest in your dataset?

The main features in the data set are platform, genres, and sales. I'm interested in determining how genre and platform performs over time and across regions.

What other features in the dataset do you think will help support your

investigation into your feature(s) of interest?

Critic and User scores may have an impact on sales, since higher scores could promote customer's to buy the game.

Did you create any new variables from existing variables in the dataset?

I created a variable User_Score that multiplies User_Score by 10 to match the format of Critic_Score. I also created 2 new columns, MeanScore which is the mean of User and Critic scores, and TotalCount which is the sum of both User and Critic counts.

Of the features you investigated, were there any unusual distributions?

Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

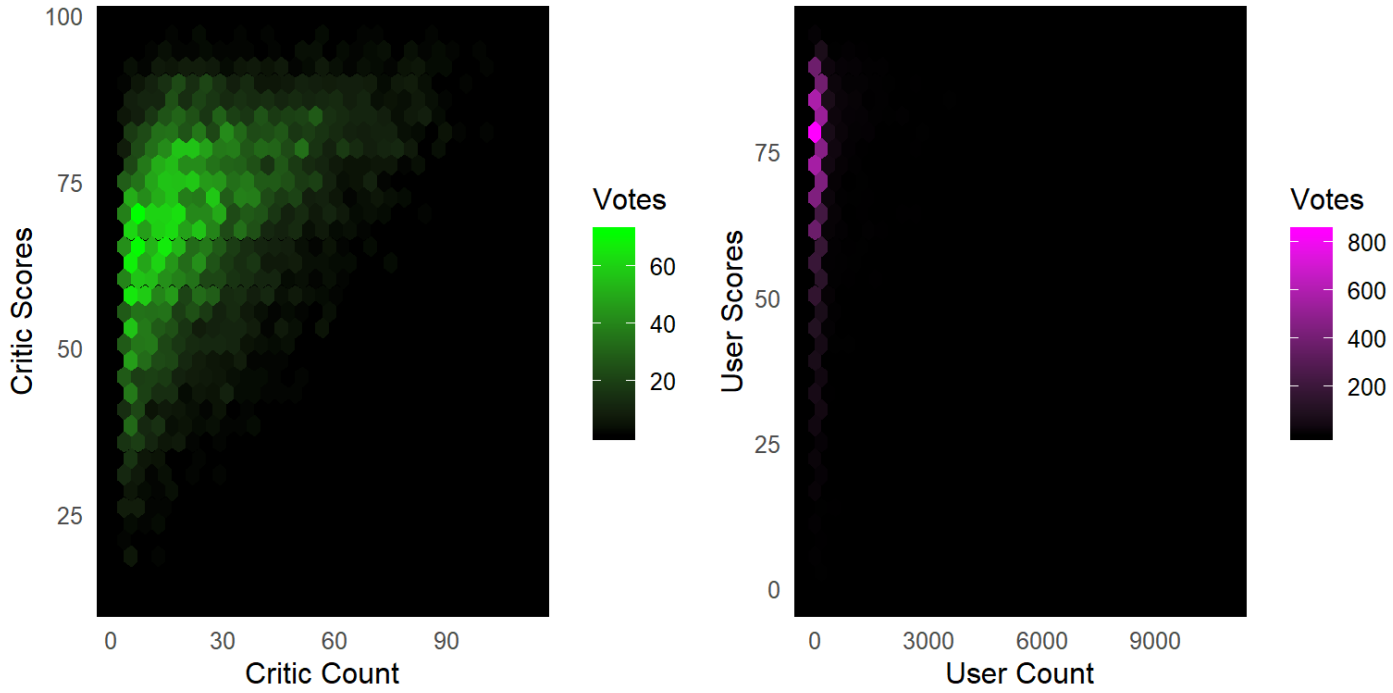
The genre and year of release distributions indicated that there were missing values in the dataset. After investigating the missing values, it was determined that 2 observations did not have a name, genre and several other features. So these observations were removed from the dataset.

It was found that Year_of_Release had multiple observations beyond the year 2016 with sales data. These observations were removed due to the dataset being compiled in Decemberof 2016. Year_of_Release was also converted to a numeric variable instead of character to allow easier plotting.

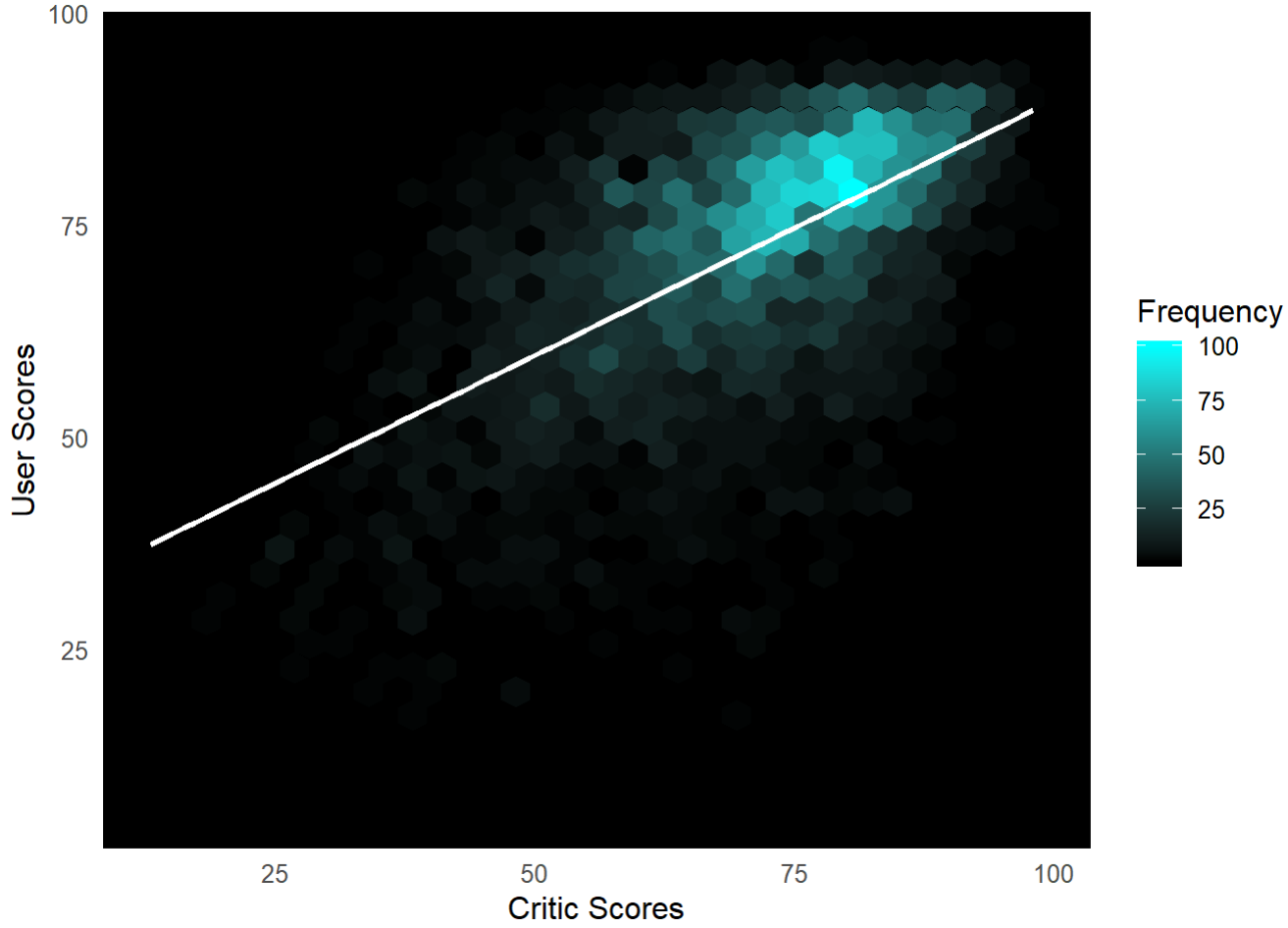
There were strings in the User Score column that were replaced with NA to prevent any calculation issues.

Bivariate Plots Section

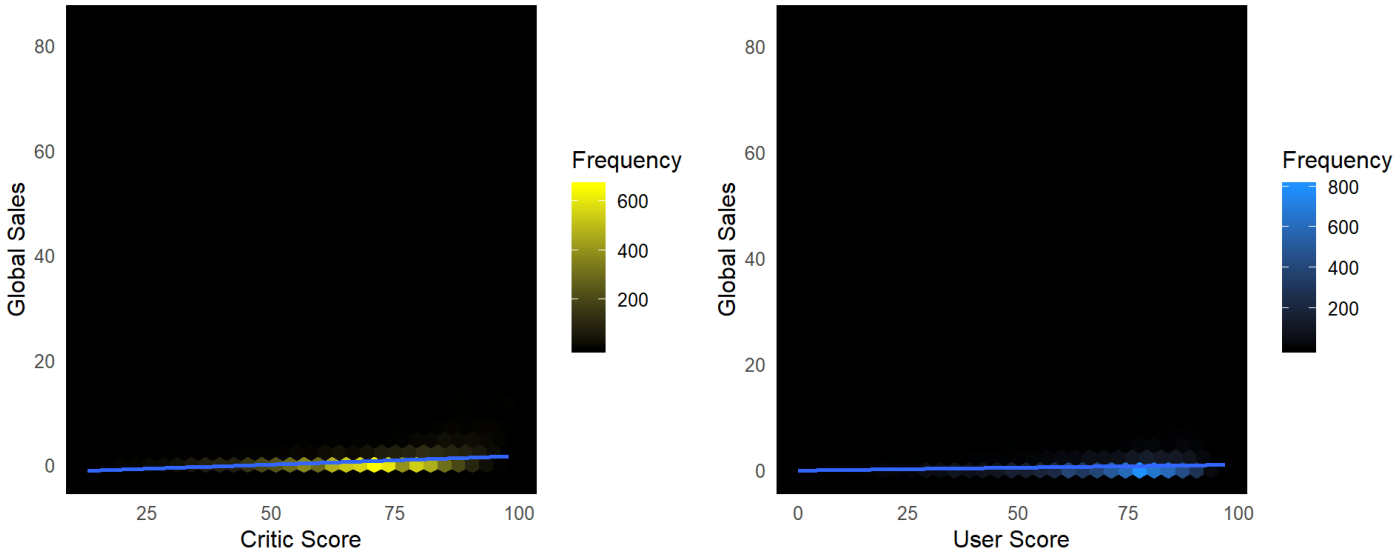
I wanted to see if there's a relationship between the reviewers and the number of reviews.



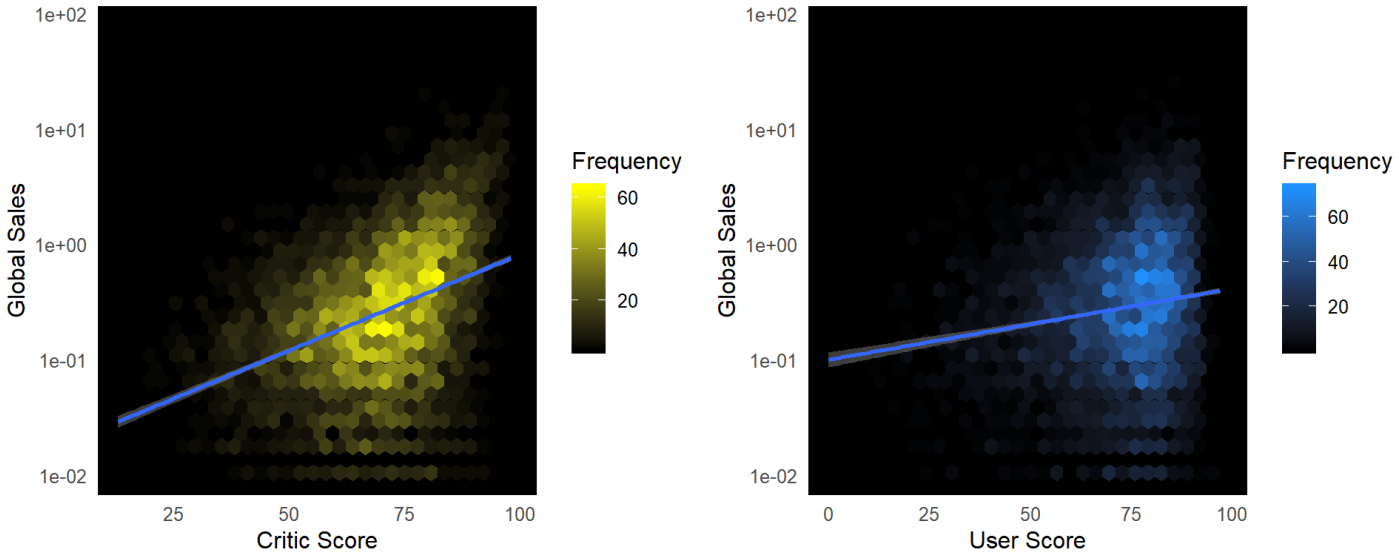
Critic Score and Critic Count seem to correlate, whereas User Score and User count doesn't seem to have any relationship. This could be due to missing values throughout the column. Does Critic and User score's have a relationship?



Critic and User Scores have a positive correlation. As critic score increases, we also see a increase in user scores. I'm interested in seeing if either scores have a relation with Global Sales.



There may be a relationship between Critic or User scores and global sales. Will transforming the global scales using a log scale help see if there's a relationship between the two variables?

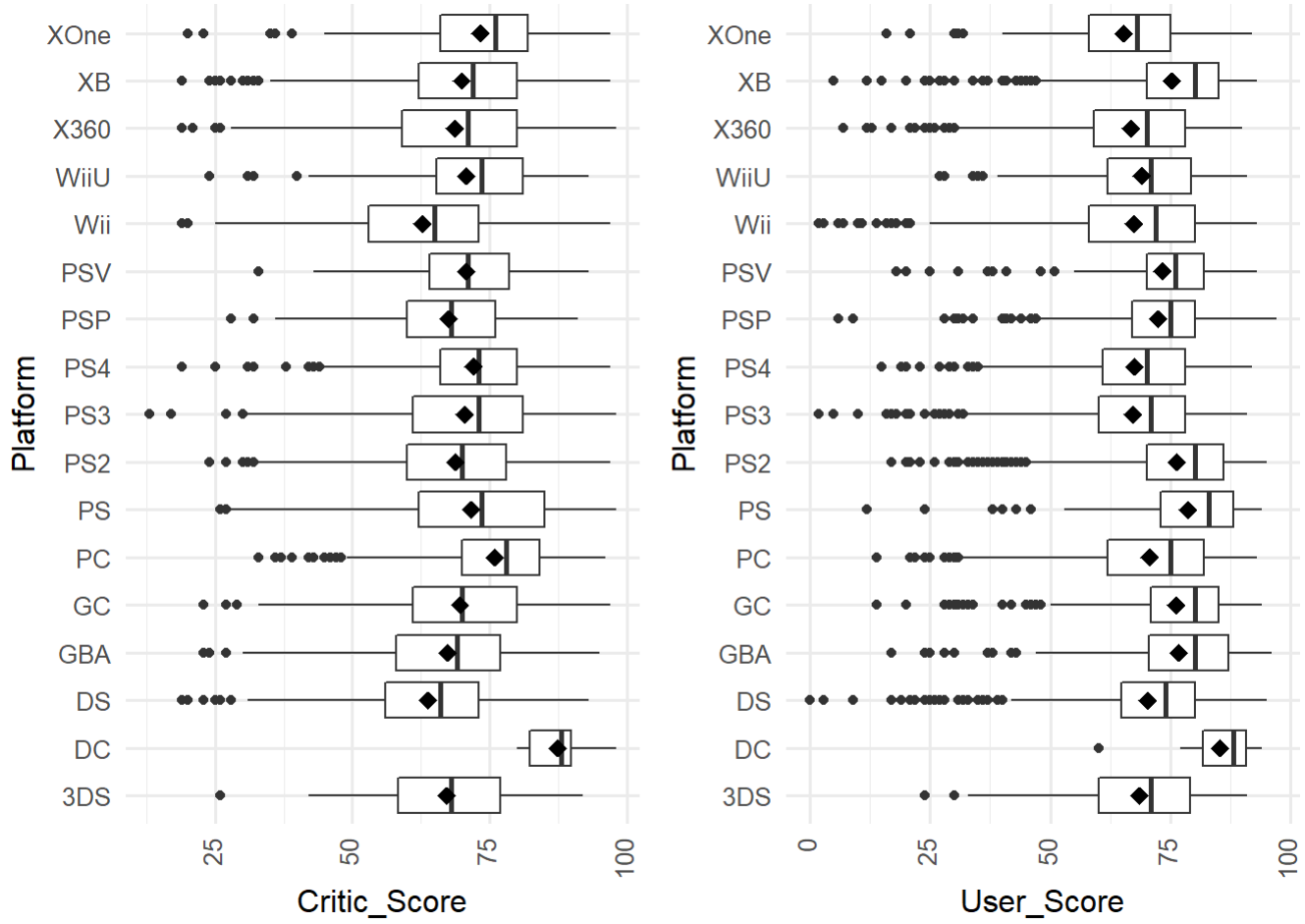



```
##
##  Pearson's product-moment correlation
##
## data:  vg$Critic_Score and vg$Global_Sales
## t = 22.607, df = 7981, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2245898 0.2658246
## sample estimates:
##       cor
## 0.2453182
```

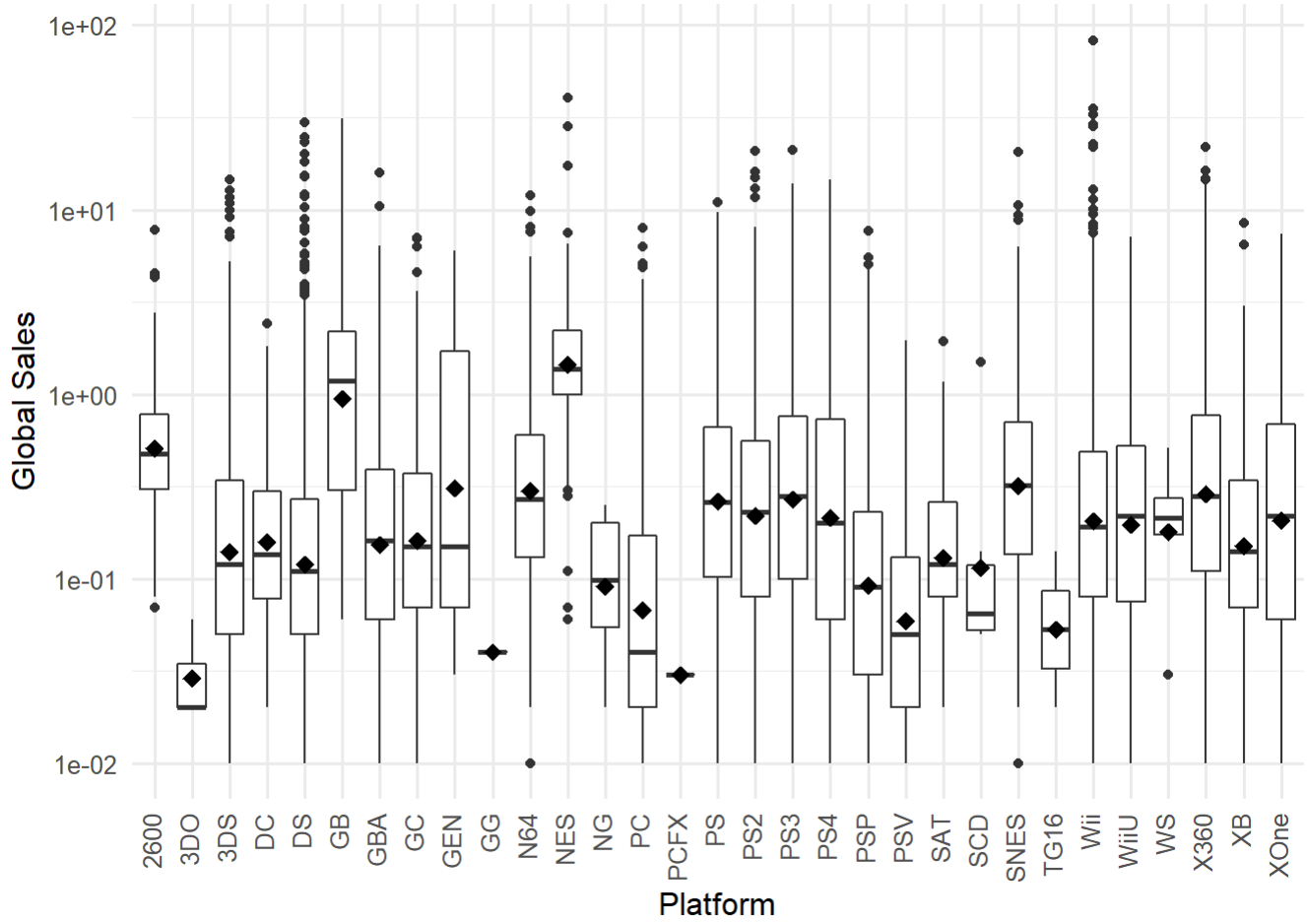
```
##
##  Pearson's product-moment correlation
##
## data:  vg$User_Score and vg$Global_Sales
## t = 7.6265, df = 7461, p-value = 2.711e-14
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.06539279 0.11041880
## sample estimates:
##        cor
## 0.08795072
```

There does seem to be a positive correlation for both Critic and Users for global sales. With Critic's having a stronger relationship with global sales than users.

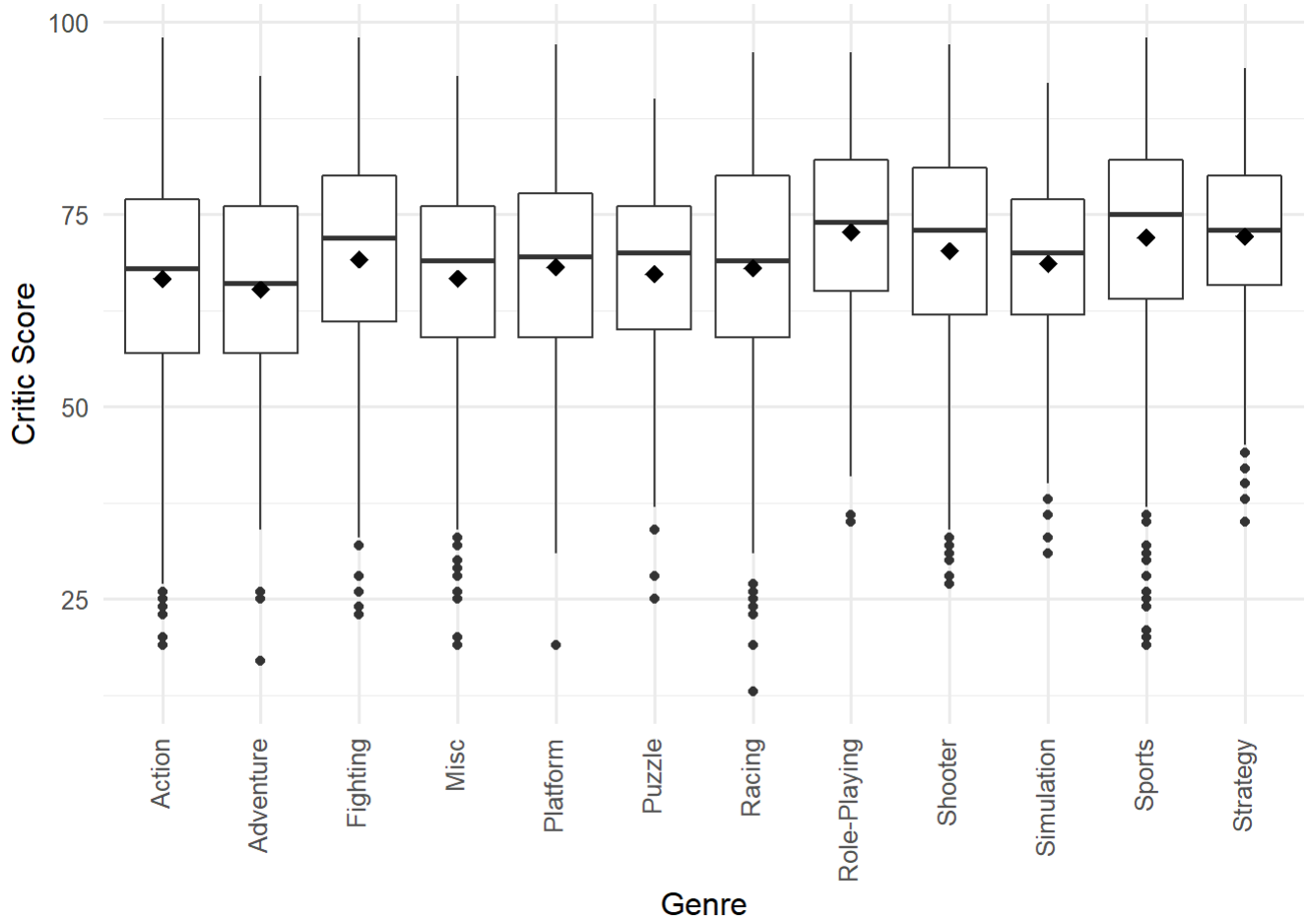
Next I'll look at the categorical features platform, genre, and year of release. I'm interesting in seeing the spread of critic and user scores across these various categorical variables.

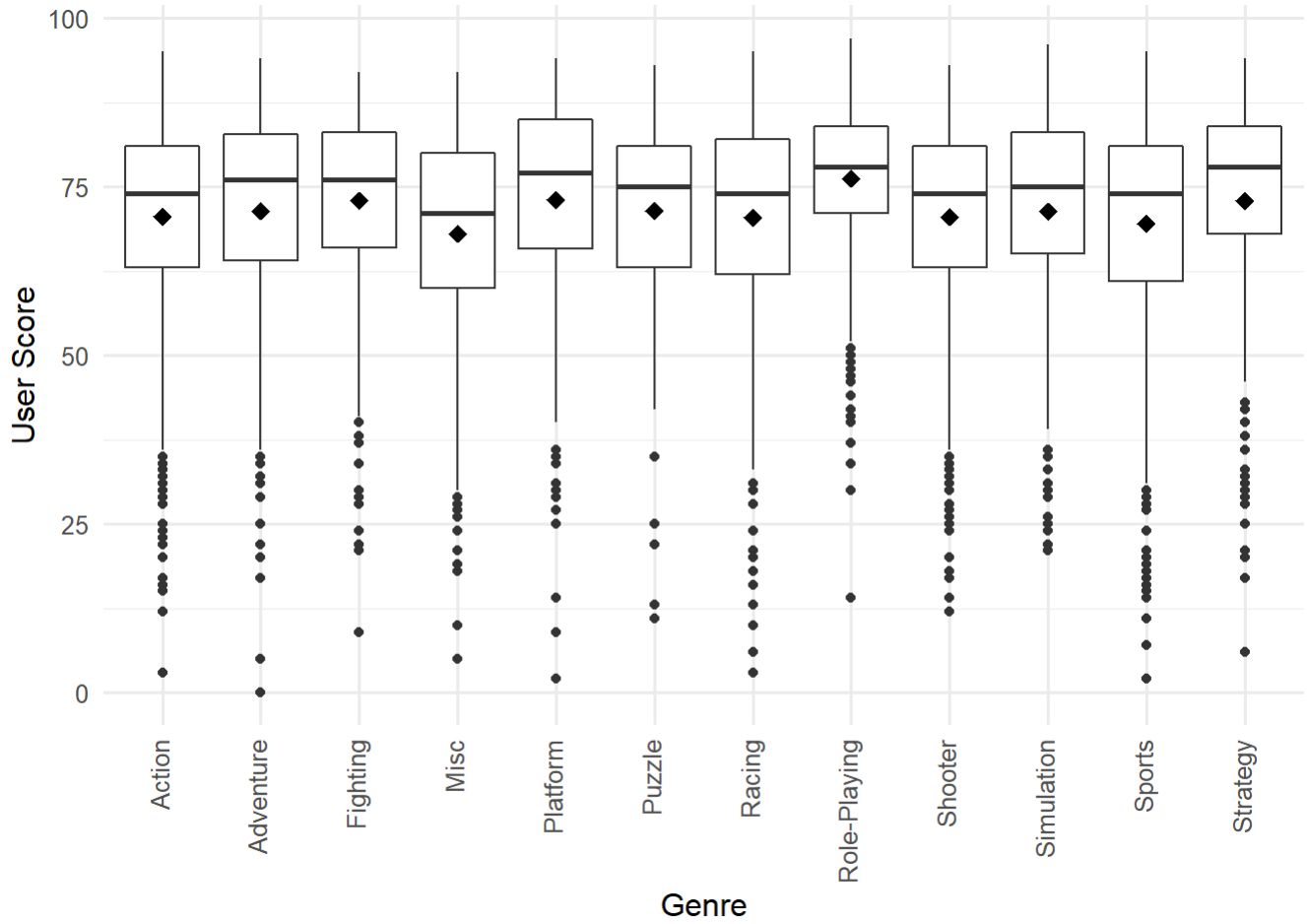


Looking at platform related observations, it becomes clear through boxplots that there are missing values when you compare User and Critic scores with global sales. There's multiple platforms that do not have either a user or critic score. It also seems the median rating a User gives a video game is higher than where critics score them. Critics also have a wider spread when it comes to scores, with user's being more close together.

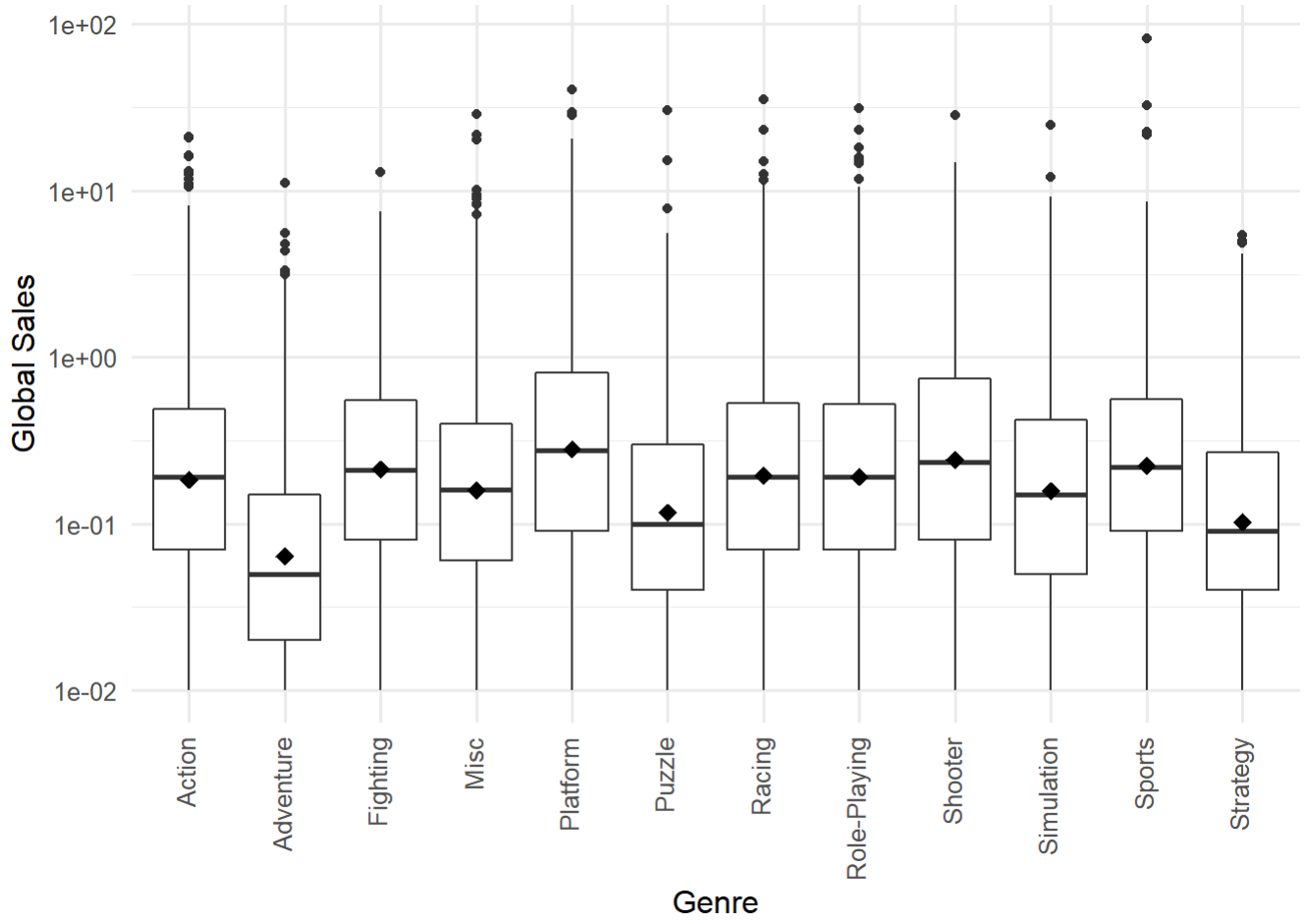


The Wii and DS seem to have a far higher number of outliers than the rest of the platforms, with the Wii having the game that's sold the most units. I wonder what genre the game is in.

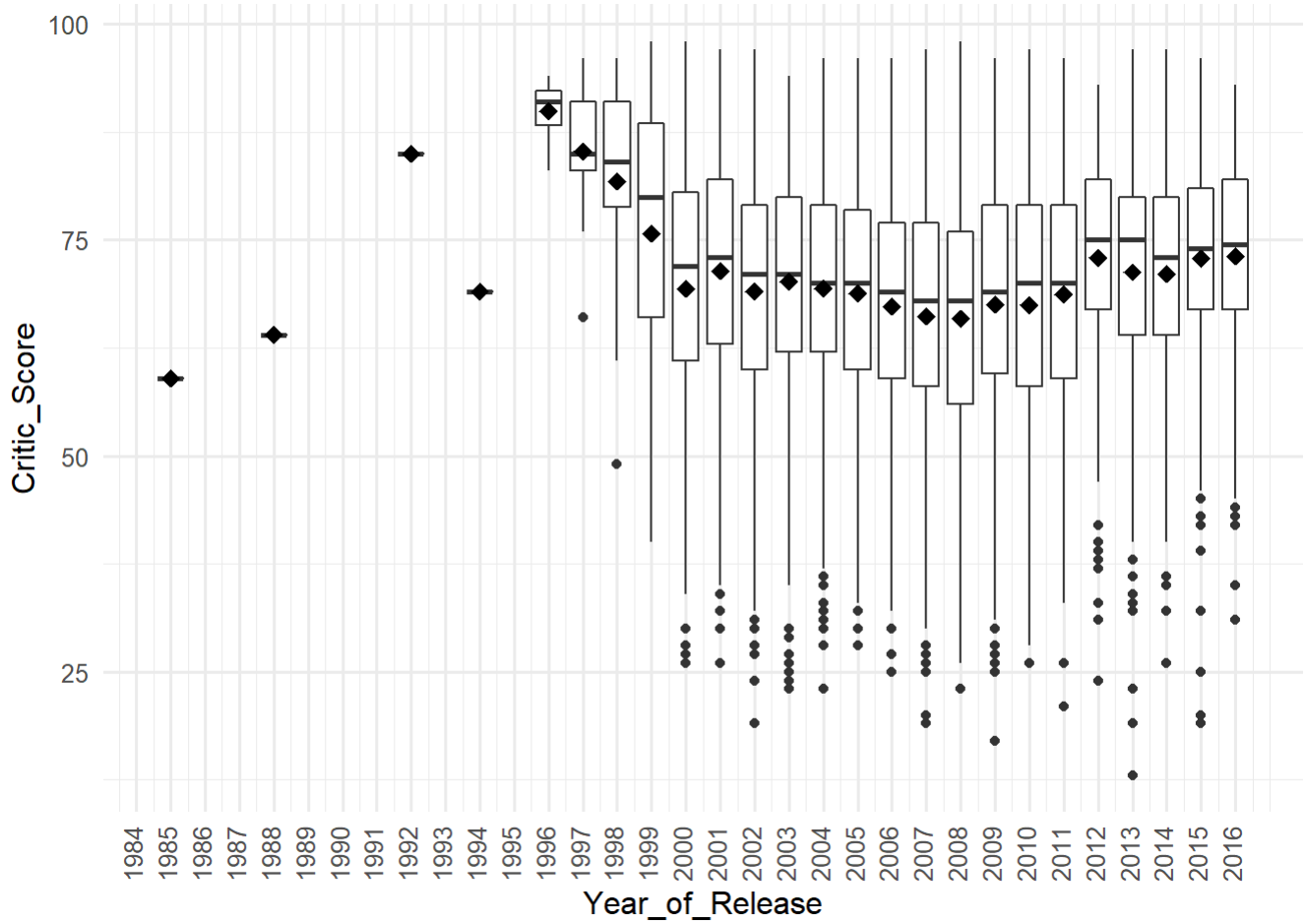

Critic Score
100
75
50
25
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role-Playing
Shooter
Simulation
Sports
Strategy
Genre

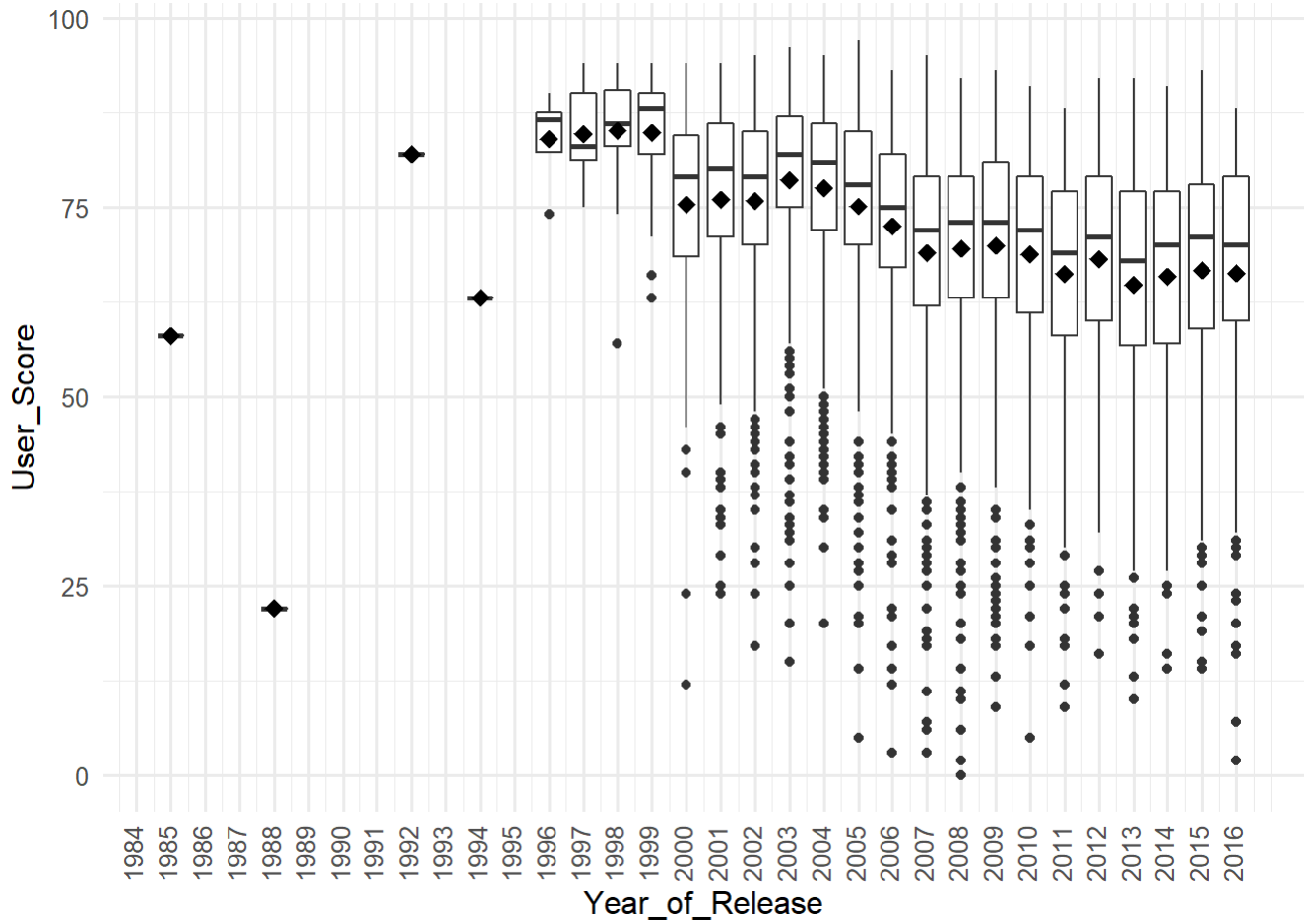

User Score
100
75
50
25
0
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role-Playing
Shooter
Simulation
Sports
Strategy
Genre

The trend of user's having a higher median rating than critics continue across genres. Users definetly rate games lower than critics across every genre. How does global sales look when we look at it by genre.



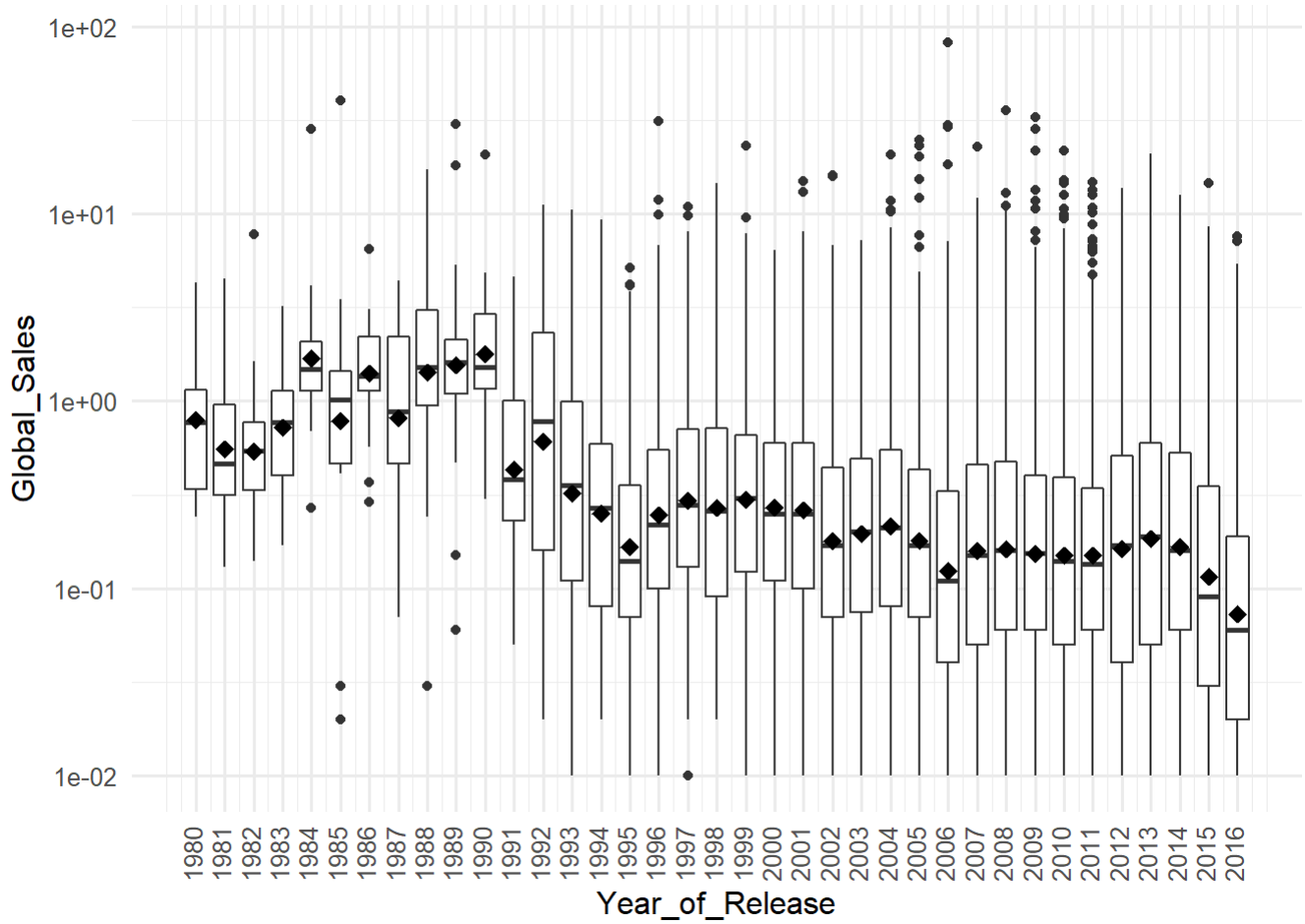
I also observed on the global_sales plot that the genre for the Wii's top selling game is most likely in the sports genre. Despite the Adventure genre having having a median user rating of over 75, it has the lowest average global sales of all genres even though it has more releases than 6 other genres (fighting, platform, puzzle, racing, simulation and strategy genres).


Critic_Score
100
75
50
25
1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016
Year_of_Release


User_Score
100
75
50
25
0
1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016
Year_of_Release

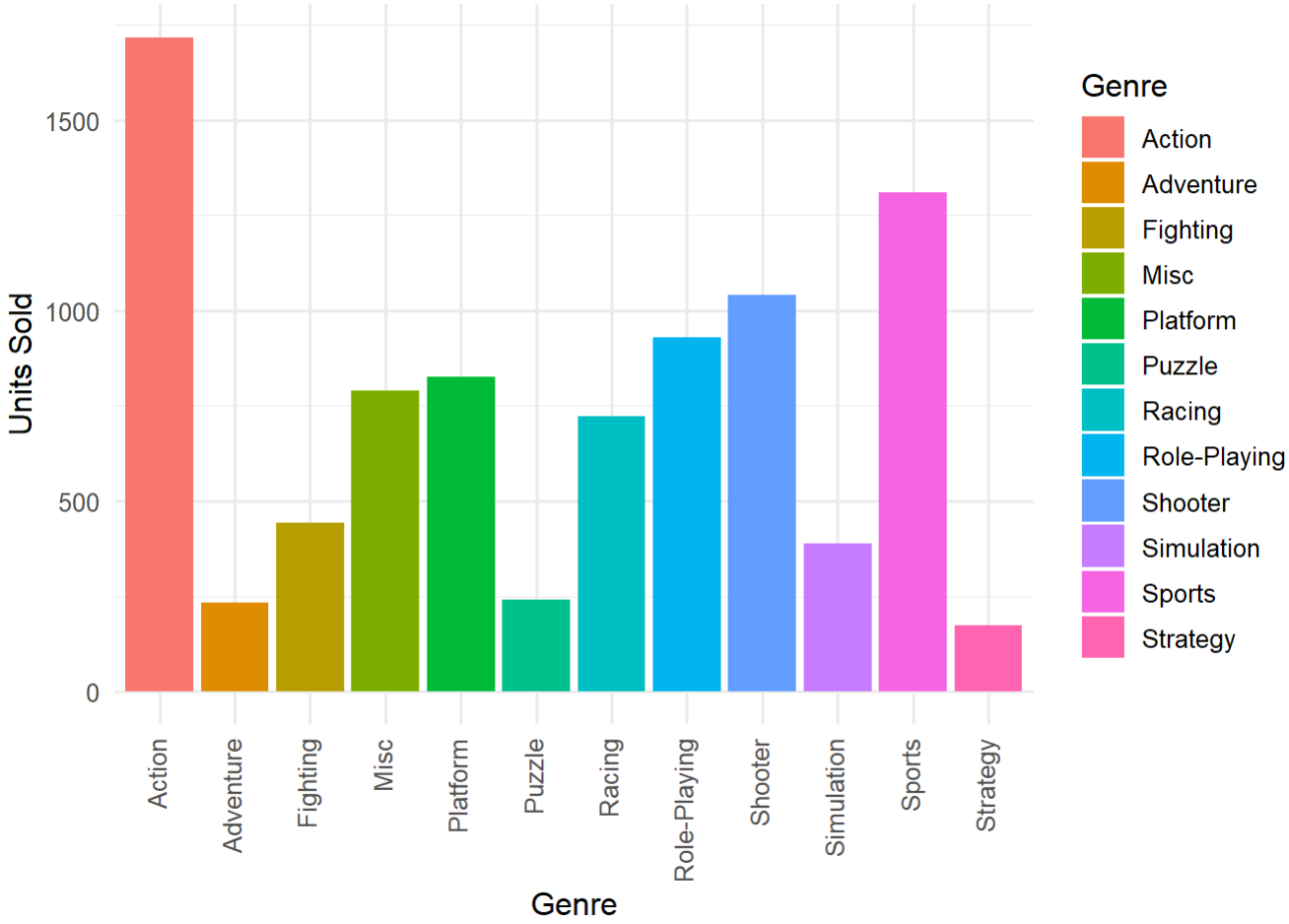
If we look at the median score for Critics the number started trending upwards around 2011 after being relatively flat for almost a decade. Critics also have a wider spread than users across almost every year. User tend to rate games more inconsistently with a large amount of outliers visible.

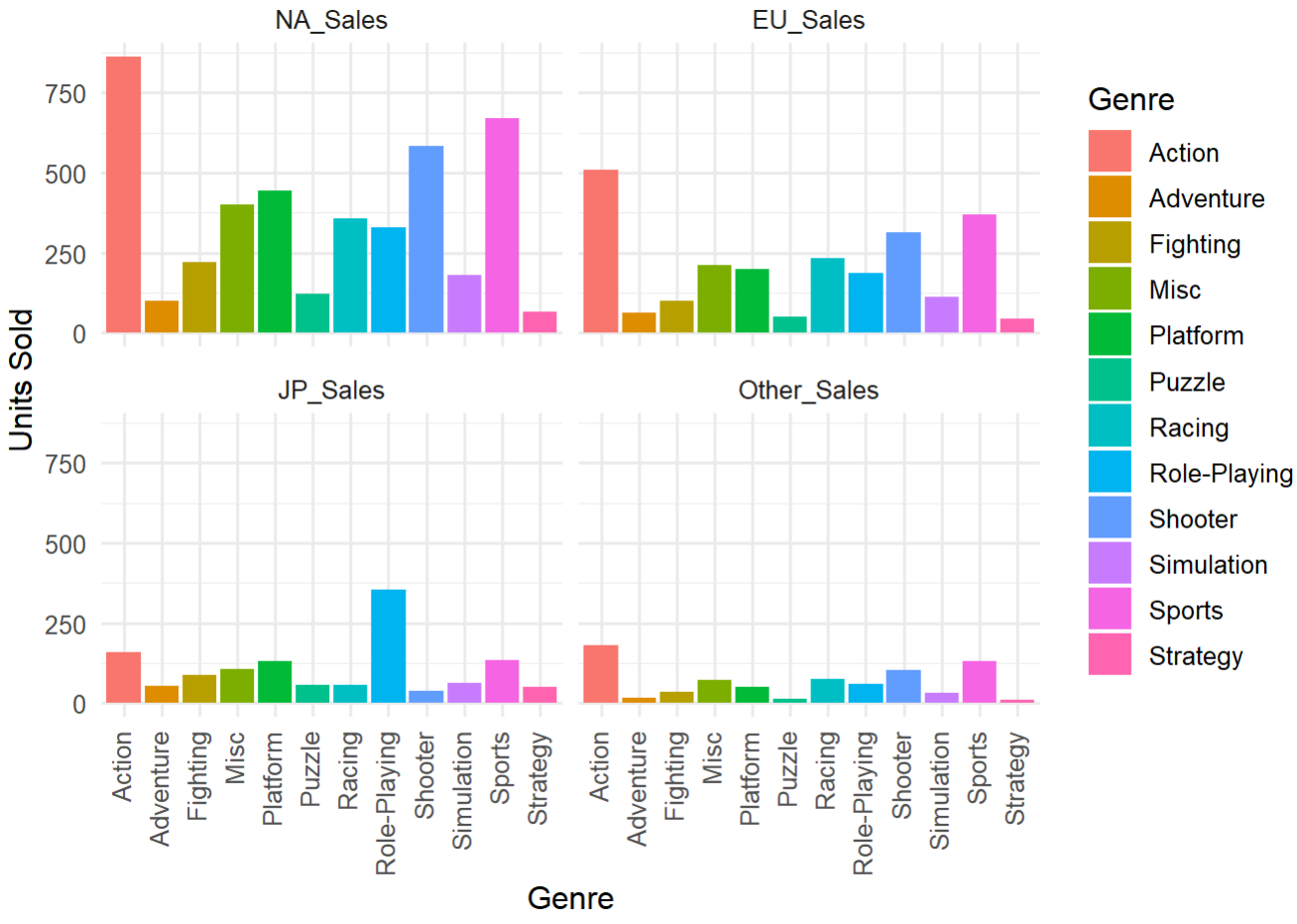
Next I'll take a look at global sales over time.



Between 2006 and 2011 Global Sales appear relatively consistent, but in 2012 the spread became wider, and continued this trend into 2016. We do see a decline in average sales starting in 2012 as well.

Next I will look at sales across regions to see if there's a genre that stands out across all regions.


Units Sold
1500
1000
500
0
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role-Playing
Shooter
Simulation
Sports
Strategy
Genre
Genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role-Playing
Shooter
Simulation
Sports
Strategy


NA_Sales
EU_Sales
JP_Sales
Other_Sales
Units Sold
750
500
250
0
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role-Playing
Shooter
Simulation
Sports
Strategy
Genre
Genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role-Playing
Shooter
Simulation
Sports
Strategy

```
## # A tibble: 12 x 2
##    Genre        Global_Sales
##    <fct>               <dbl>
##  1 Action             1718.
##  2 Sports             1310.
##  3 Shooter            1042.
##  4 Role-Playing        931.
##  5 Platform            826.
##  6 Misc                791.
##  7 Racing              724.
##  8 Fighting            443.
##  9 Simulation          388.
## 10 Puzzle              240.
## 11 Adventure           233.
## 12 Strategy            173.
```

If we look at Global Sales by genre we can see that the Action genre significantly outsells the other genres. I was curious if this trend was common across all regions, and it turns out that it is,with the exception of Japan which has more sales from the role-playing genre than the action genre. North America clearly outsells the other regions in almost every genre.

Bivariate Analysis

Talk about some of the relationships you observed in this part of the

investigation. How did the feature(s) of interest vary with other features in
the dataset?

Some of the relationship's observed is with the way Critic Scores and Count correlate with each other, but there doesn't seem to be a correlation between User Scores and Counts. However, there's a positive relationship between User and Critic scores. They both commonly rate games around a 75.

Did you observe any interesting relationships between the other features

(not the main feature(s) of interest)?

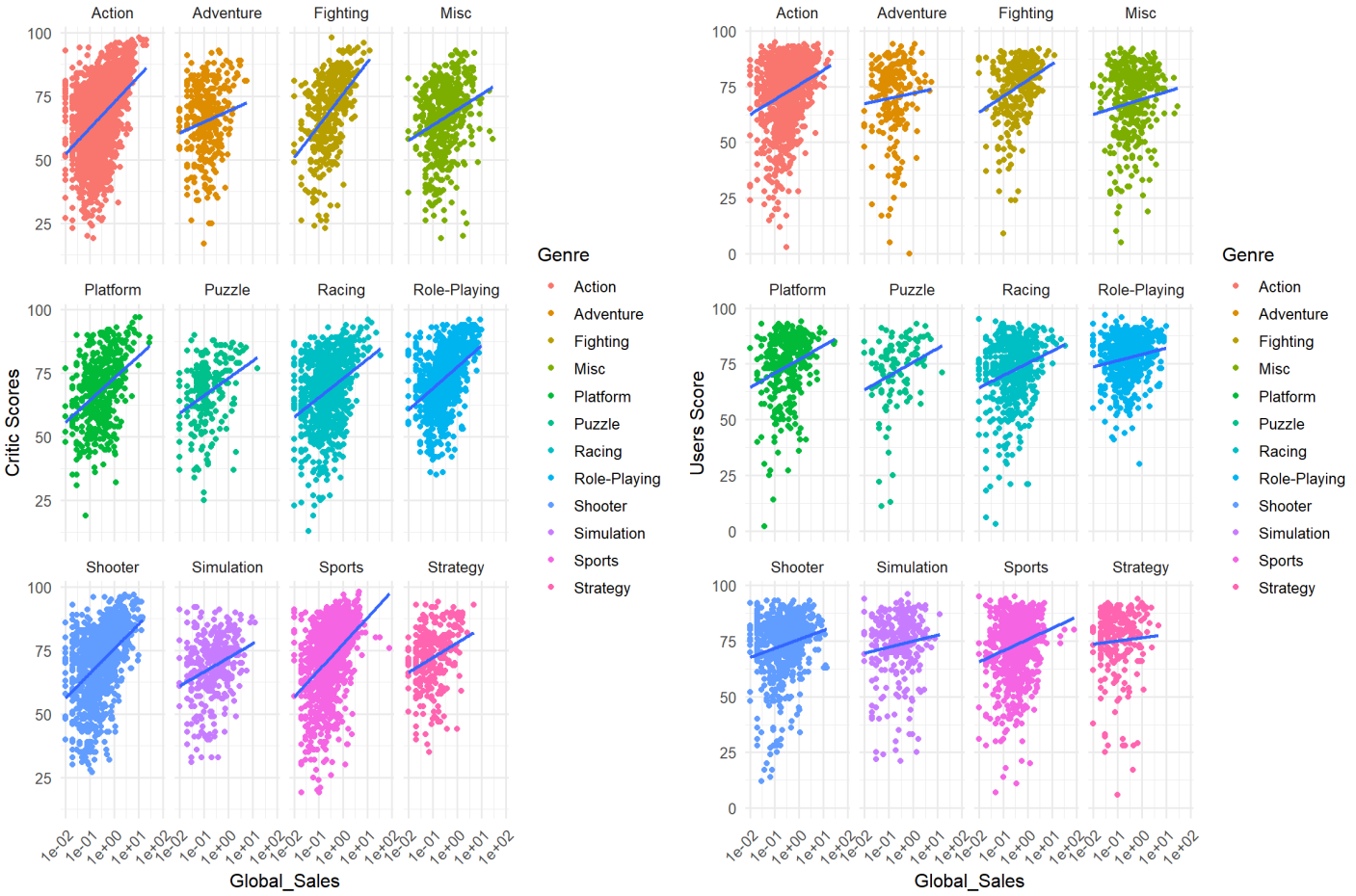
I was expecting to see a positive relationship based off of the hexbin plots between User/Critic Scores and Global sales. However, when you look at the boxplots there's a significant amount outliers for poorly rated games, and but there's multiple outliers across global sales.

What was the strongest relationship you found?

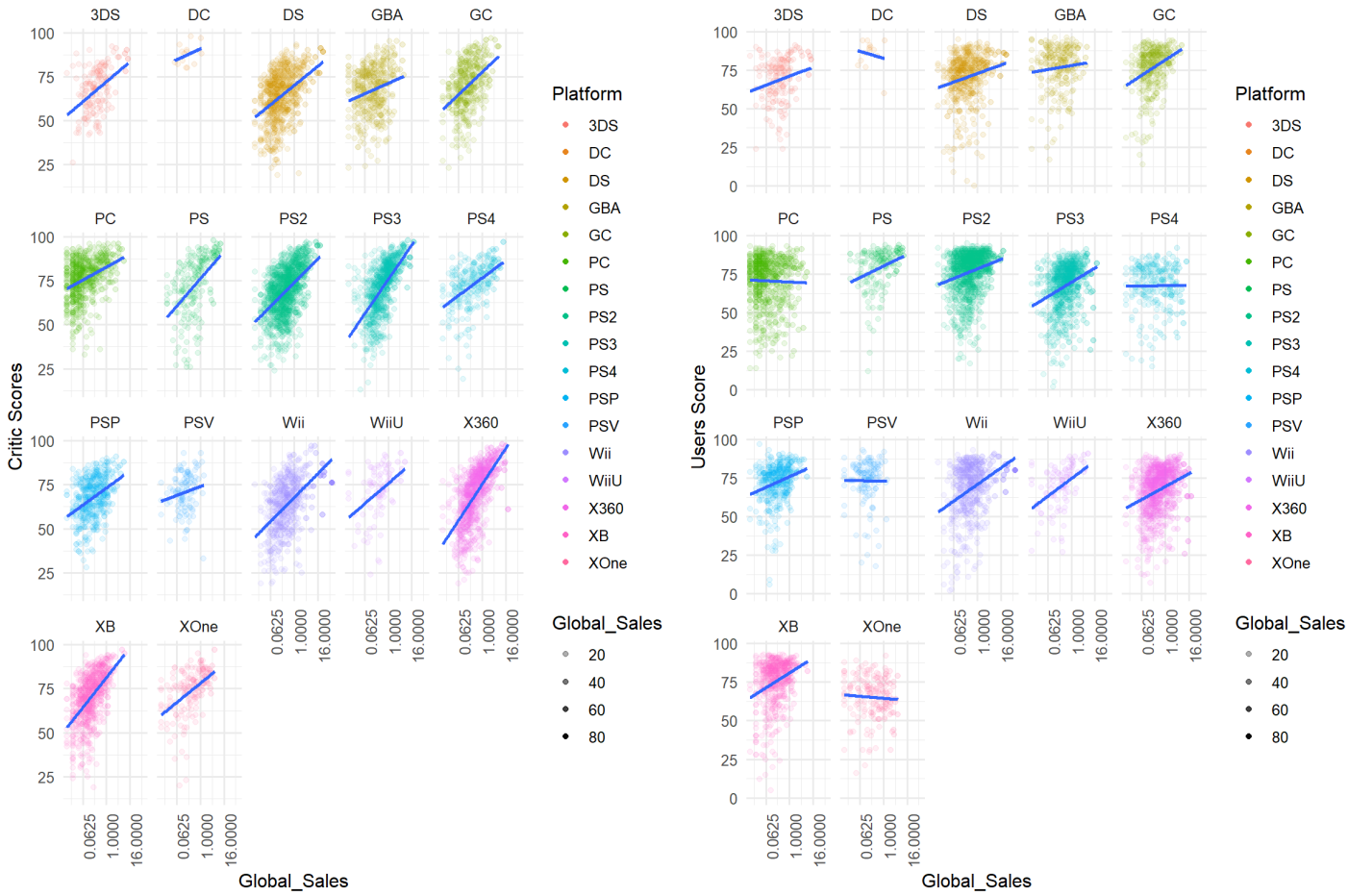
The strongest relationship observed would be between genre and sales. Even though the action genre has an average score below 75 across Critics and Users, it accounts for the highest amount of sales across every region except Japan.

Multivariate Plots Section

Next I'll take a look at multiple variables to analyze relationships further.

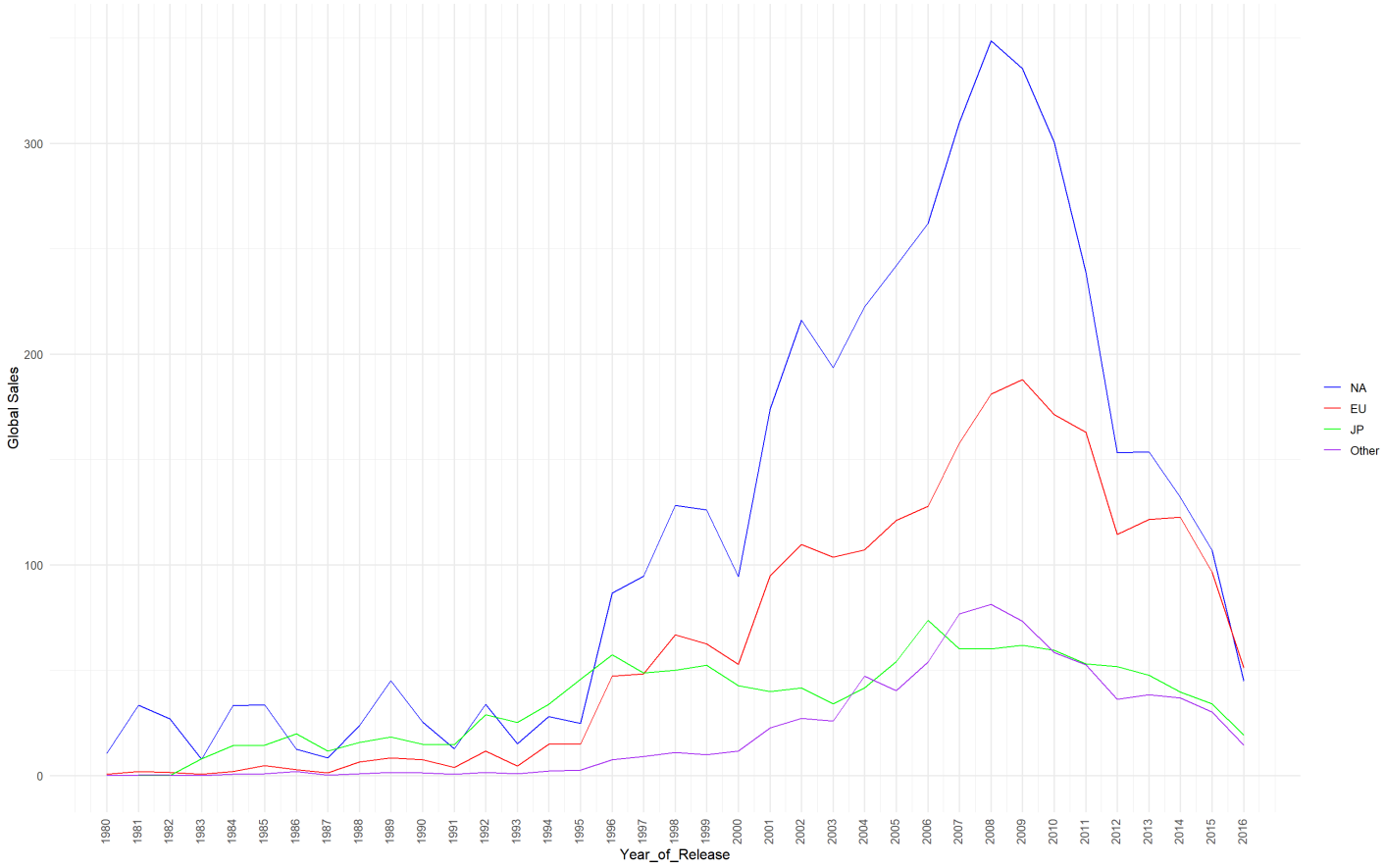


Across almost every genre, critics has more of a positive correlation than users. It also seems like critics have a wider spread of review scores which matches with what I observed in the boxplots.



I was thrown off by this plot initially. I was not expecting multiple platforms to have a negative correlation with user scores and global sales. Critics doesn't demonstrate a single negative correlation with platform and global sales.

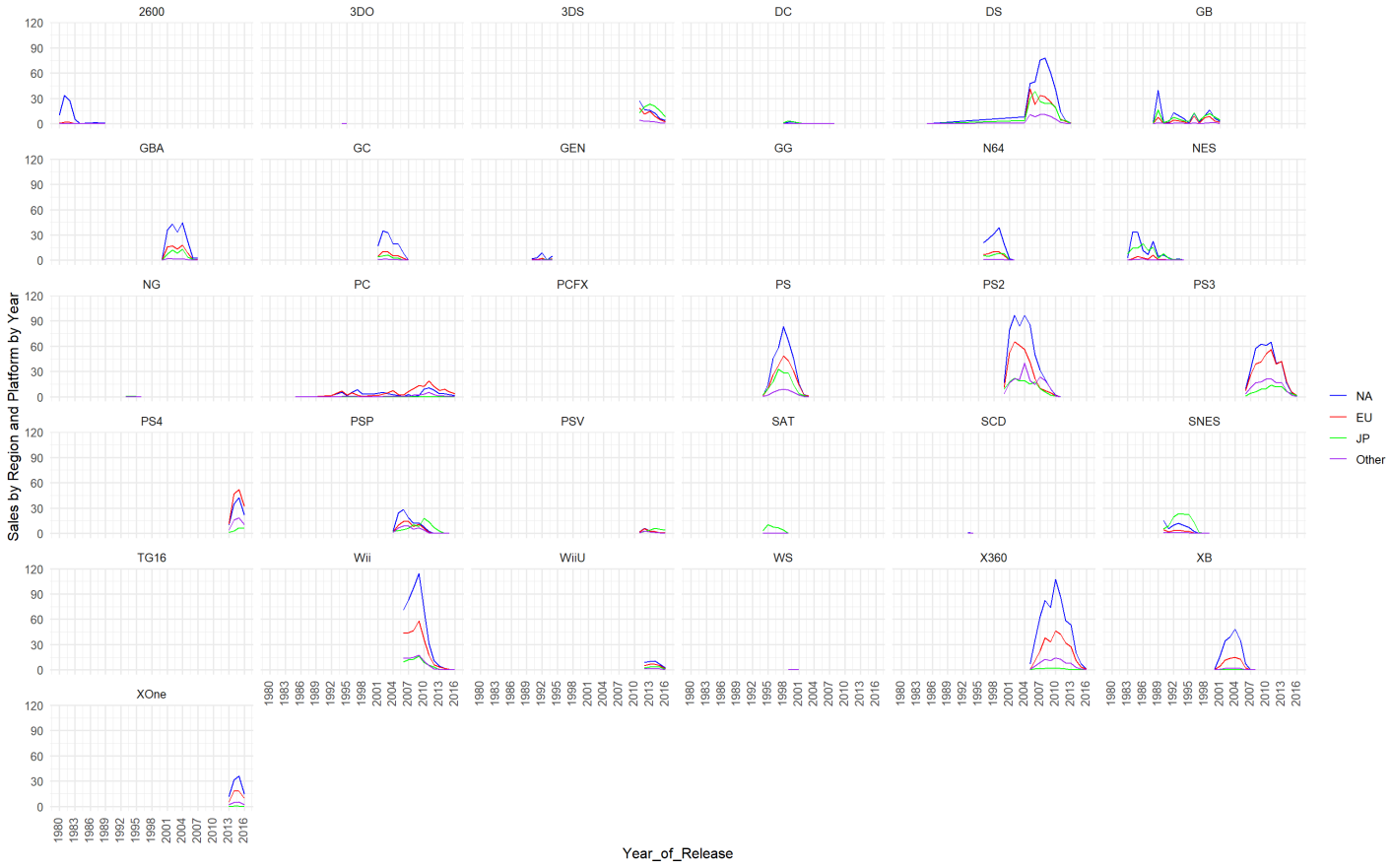
Next I wanted to see the trend of sales over time by Region.



Almost since the creation of home entertainment systems, NA has lead the global sales with only a few exceptions. I never expected NA to outsell every region so dominately after the year 2000. NA maintained this lead until 2016, when the entire global market decreased. Could this be a fault in the data?

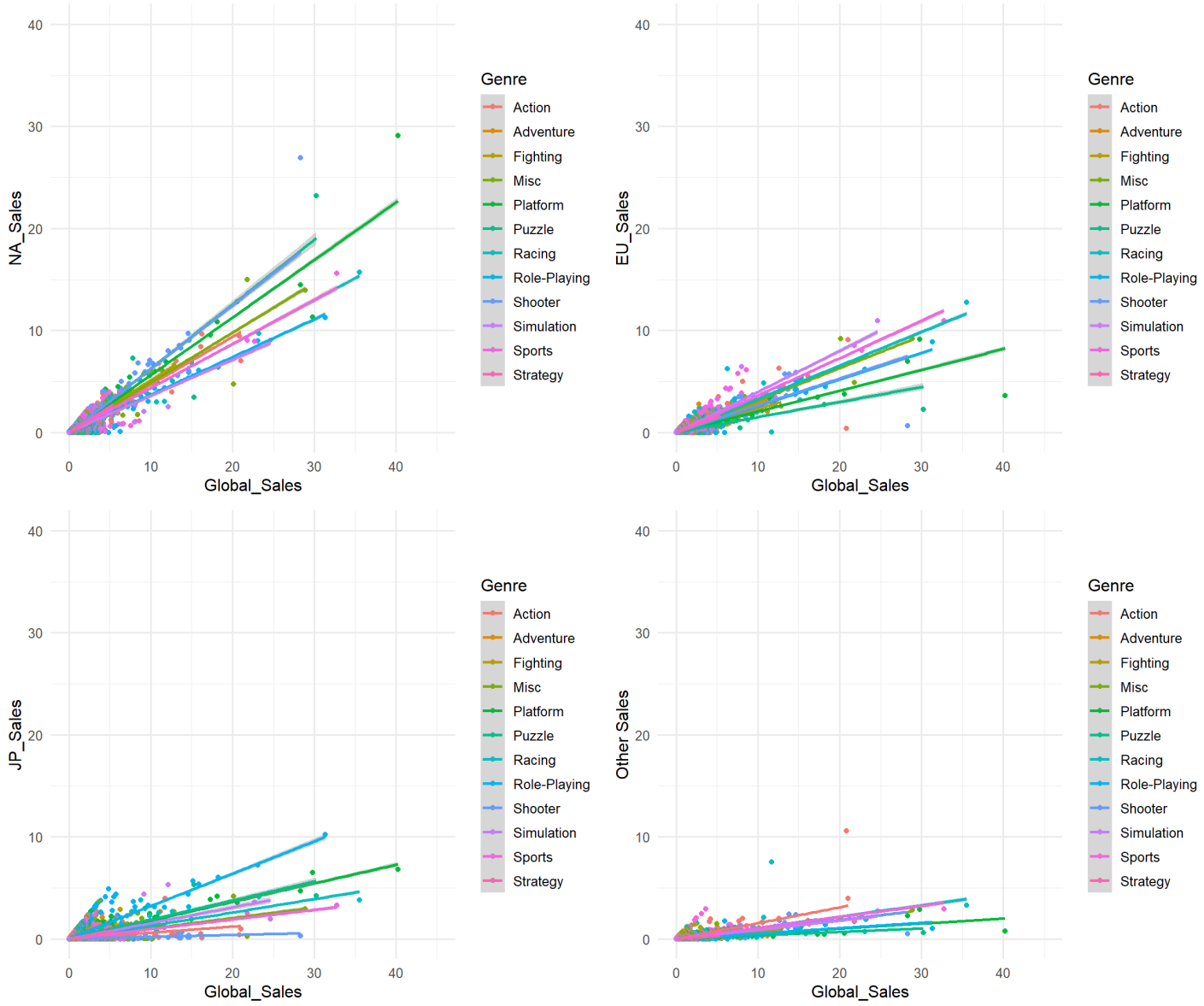
Here I can see that at the start of the video game era, the platform genre accounted for a significant amount of the global sales, and slowly lost it's appeal based on sale figures. Action and Sports games became the dominate genres starting in the late 90s.

By region, the same trend emerges except for the JP region. Role-playing games have consistently been the highest contributor to sales, with the action genre overtaking it starting in 2014.

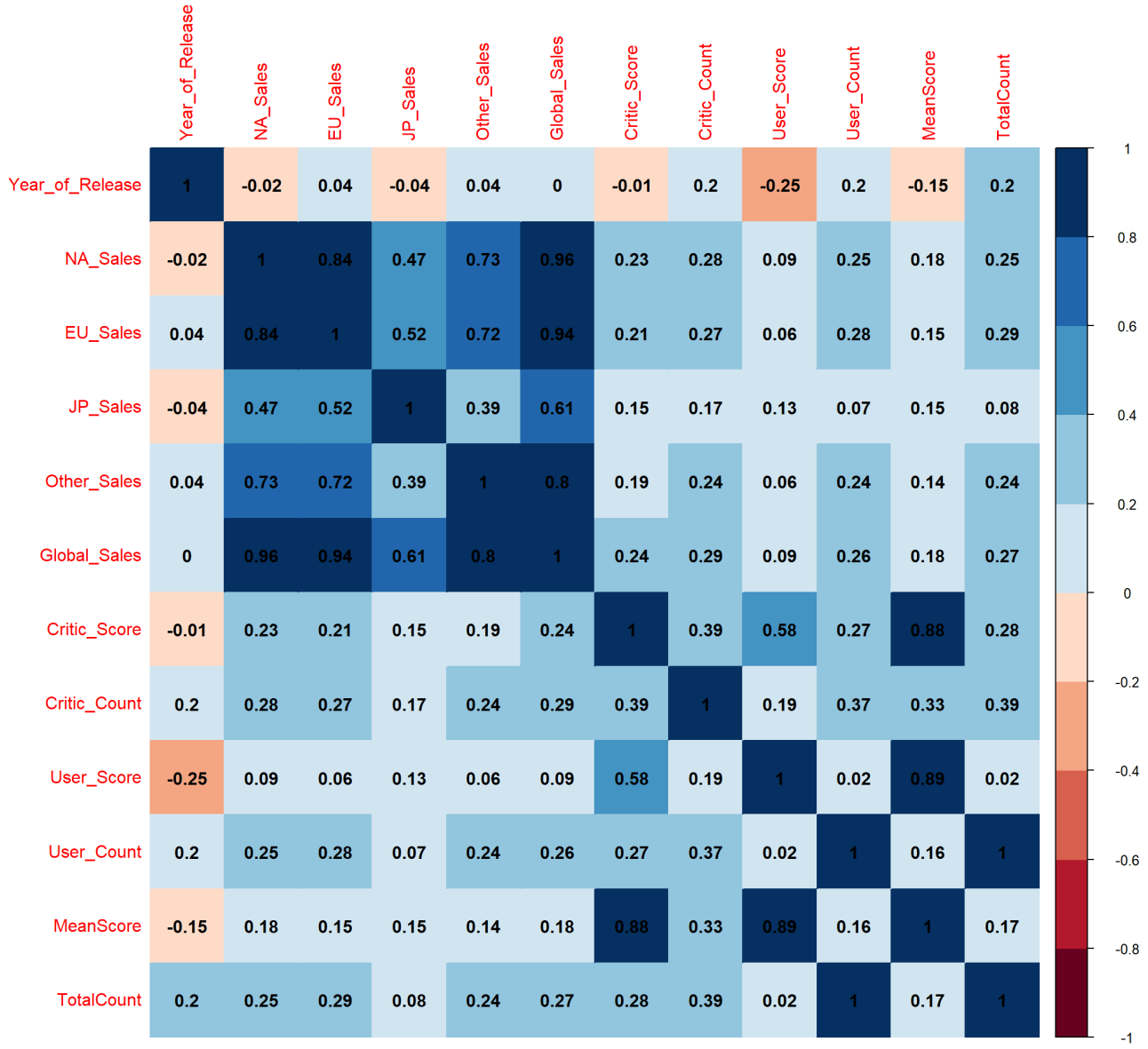


Plotting platforms over time, you can almost pinpoint exactly when a specific platform was released and reached end of life. I was impressed with how long the GB was on the market capturing sales. I was surprised that a few regions outsold NA, the PS4, SNES, SAT and PSP all sold better in other regions for a time period.

Next I wanted to take another look at the potential relationship between scores and global sales.



It's clear that NA_Sales has a strong positive relationship with Global Sales. For NA I was expecting either action or sports to have the strongest correlation with Global Sales, but it actually looks like the Shooter genre has the largest positive correlation.



NA_Sales has the strong correlation with Global Sales out of all the other regions here, which was to be expected. Critic scores have a much stronger positive correlation to all sales regions than users. The variable that caught my eye was User Count, and Critic count which both actually have a more positive relationship than the actual scores does on sales.

Multivariate Analysis

Talk about some of the relationships you observed in this part of the

investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

Genre has an impact on sales, the Action and Sports genres for example outsold every other genre while not having the best scores. All the regions share a relationship with Global Sales as well, with NA_Sales having the strongest relationship amongst the regions.

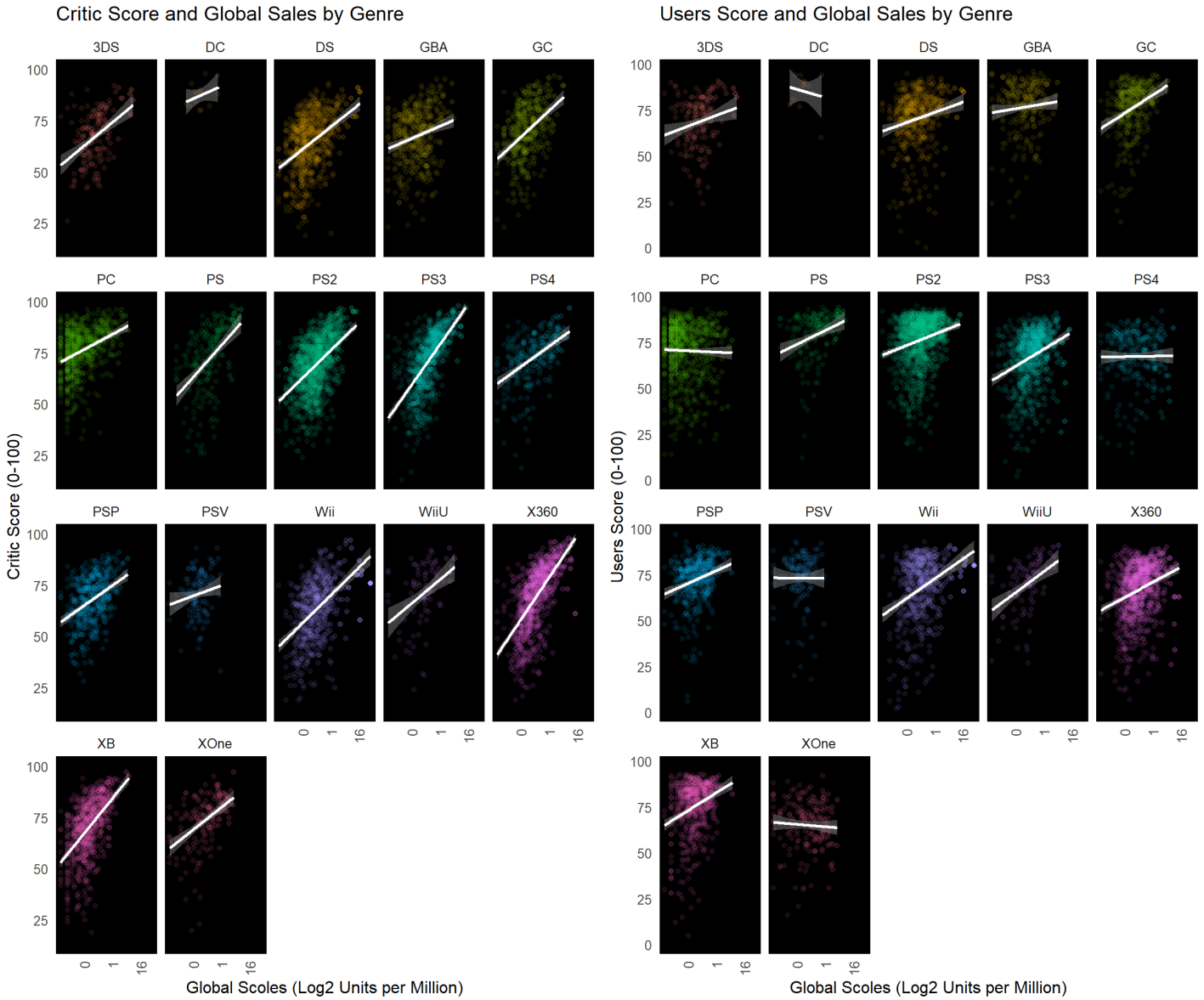
Were there any interesting or surprising interactions between features?

I was surprised at the interaction between user scores, global sales and the individual platforms. I was not expecting several platforms to have a negative correlation with user scores and global sales.

The other interesting interactions are the user and critic count relationships with sales. Almost every sales region have a more positive correlation with the counts instead of the scores. This could be due to missing values in the scores, but it intrigued me.

Final Plots and Summary

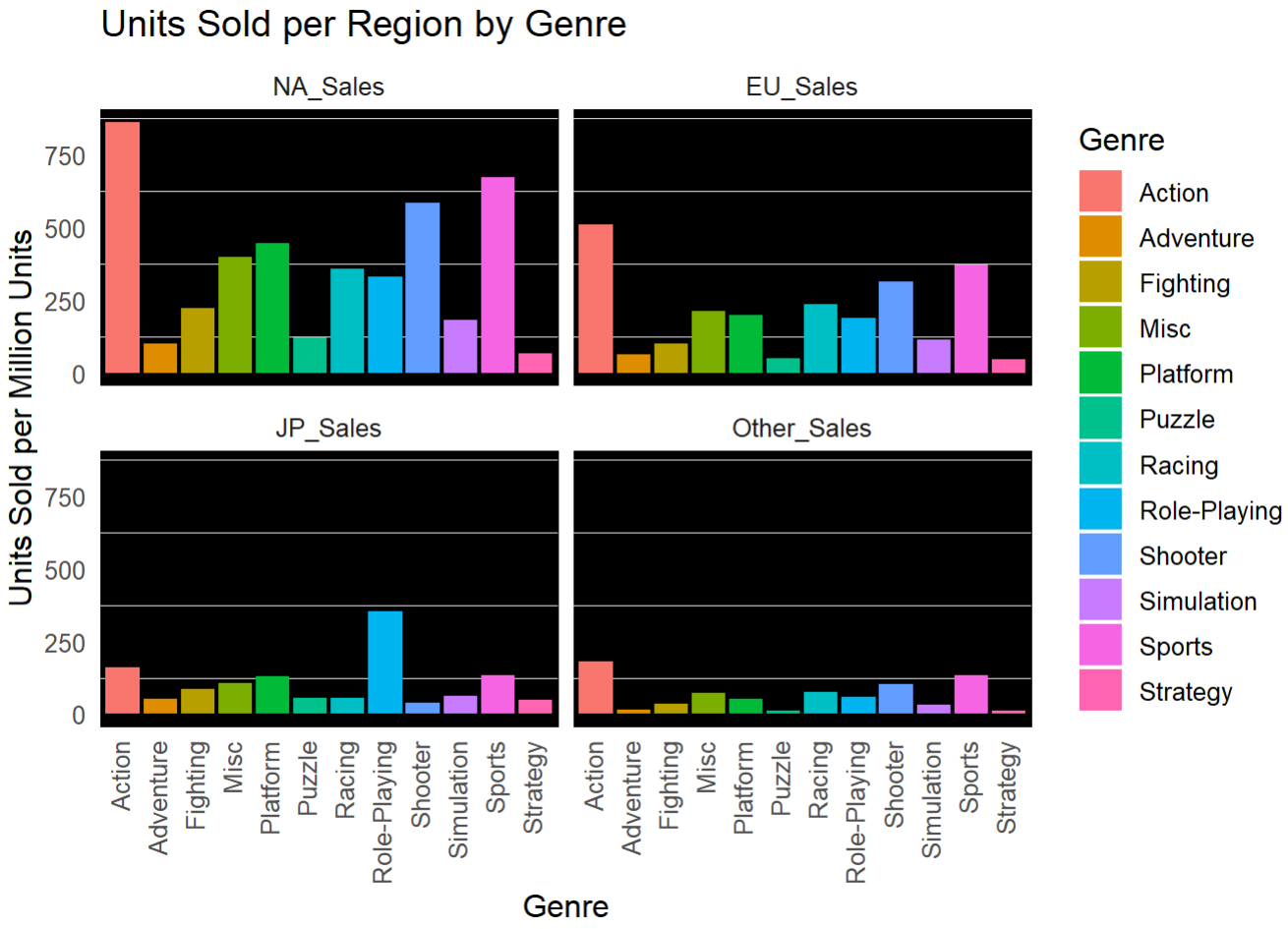
Plot One



Description One

I select this plot because it indicates a difference between the users and critics across platforms. It illustrates how critics have a positive correlation across the board for every platform. However, users demonstrate a negative correlation across 5 platforms.

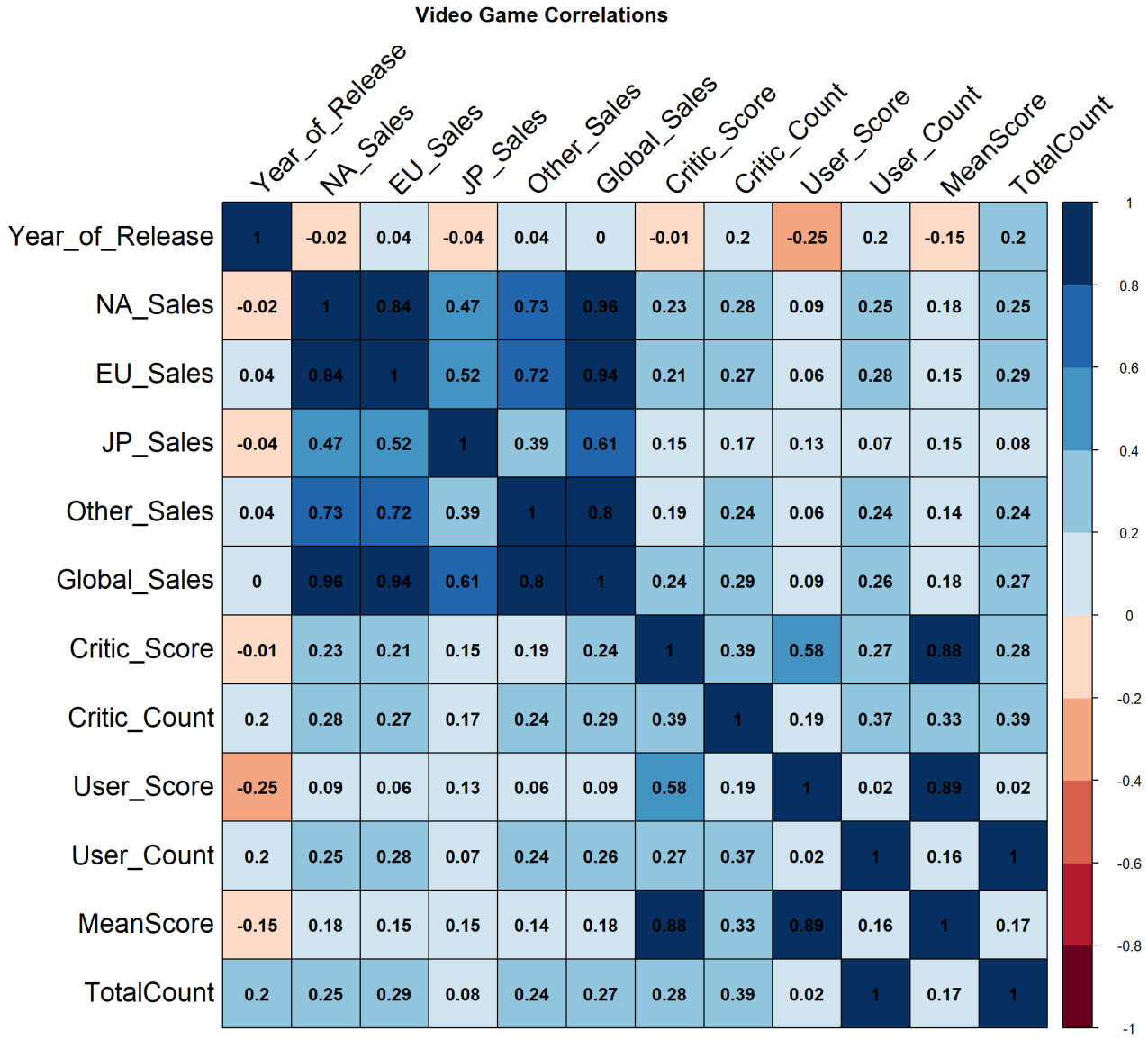
Plot Two



Description Two

This plot was chosen because it indicates very clearly that the Action genre is the #1 seller in all regions except Japan. It also shows that the sports genre is the 2nd most popular, once again with the exception of Japan. In Japan, Role-Playing games outsell every other genre significantly.

Plot Three



Description Three

The coorelation plot was chosen as the final plot, because it shed some light in an area that I did not see in other plots. The relationship between user and critic counts have more of a positive correlation than the action scores in almost every instance, and in some instances drastically different. The only exception is once again Japan, where we don't see the same relationship for users. The plot organizes information very well, and you can quickly identify the potential relationships between variables.

Reflection

The videogame data set contains information on approximately 16, 700 observations across 16 variables from 1980 to 2016, and it was collected from a single source website Metacritic. I started by understanding the various variables within the dataset. I explored questions I had about the dataset, and continued to make observations from the various plots.

There was success during my analysis of the dataset. There's a correlation between specific genre types, platforms, and critic and user scores to a less extent. For example, I was quite surprised at exactly how many sales the Action genre had across every region, as well as sports. Japan having higher sales for the role playing genre is also curious. I wasn't expecting this genre to signifcantly outsell the other genres as much as it did. The correlation that surprised me the most, has to be the results from observing how user and critic counts correlate to sales. I did not expect them to have a more positive correlation than the actual votes. However, I saw with the exception of Japan that counts were either more positive or within a few points of the actual scores.

The limitations I encountered with the dataset includes the sorce data. There's a significant amount of data missing, along with only coming from a single source Metacritic. The website launched in 1999, and has data going back to the 1980. There are several observations that do not have metrics available for them, more commonly among the user, critic, publisher and rating variables. These missing values could have impacted the observations I've made significantly, and ultimately is why I did not use the MeanScore and TotalCount more throughout the analysis. To investigate the dataset further, I would prefer to validate the data against multiple other sources to have a more thorough and validated collection of observations. This could lead to further insights and potentially help identify additional relationships between the variables.