Sales of Video Games Exploration by Tony Nguyen

Summary

This report is about exploring video game sales throughout the years. It should be noted that sales in this report refer to units sold not profit generated. My goal is to get a general understanding of the history of the sales, and how world events and trends could possibly affect the sales data.

The data was found through a website called VG Chartz and credits for extracting and cleaning the data goes to Kaggle user GregorySmith.

Kaggle url - https://www.kaggle.com/gregorut/videogamesales (https://www.kaggle.com/gregorut/videogamesales) VG Chartz - http://www.vgchartz.com/ (http://www.vgchartz.com/)

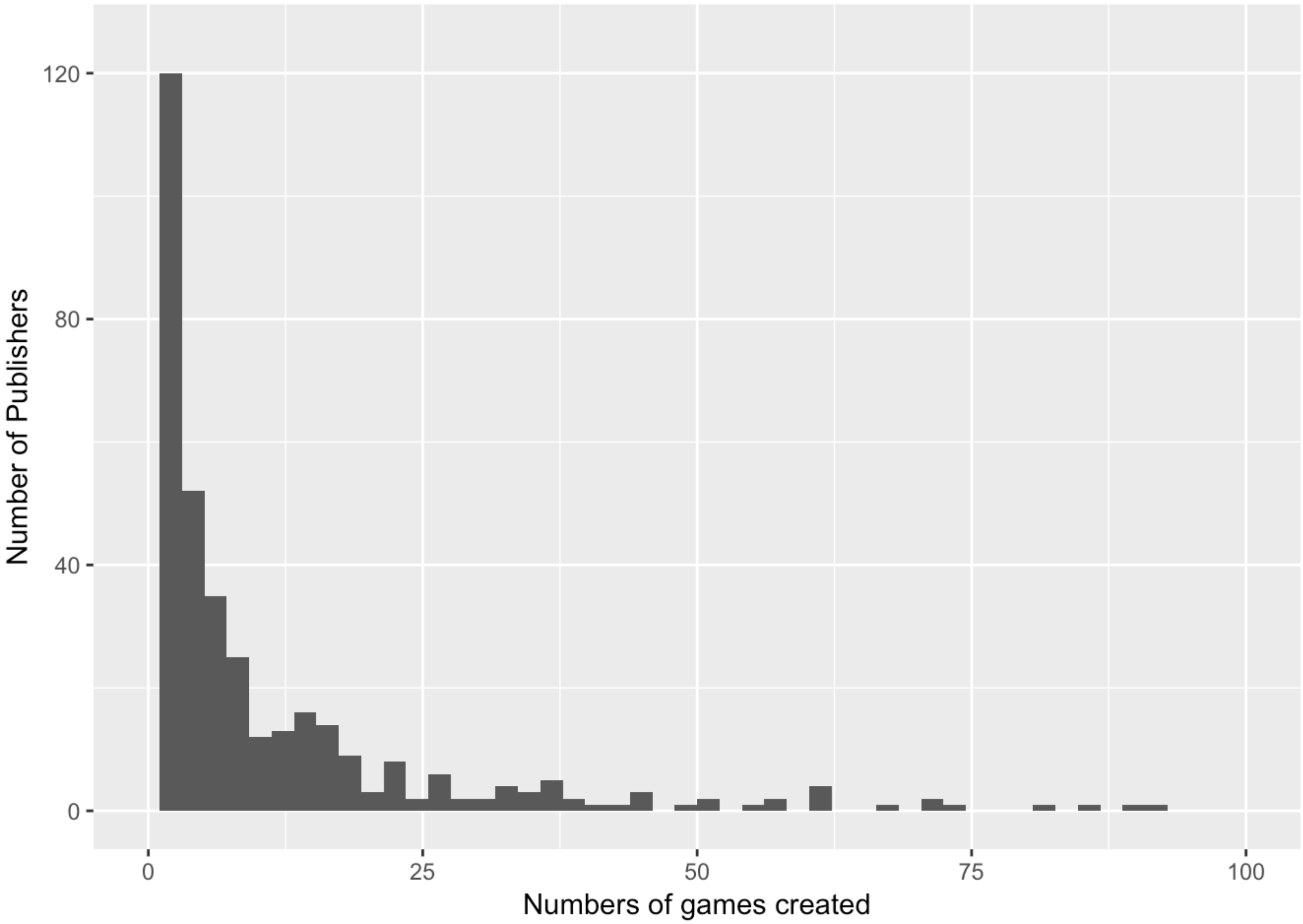
I want to empathize that I did NOT look at the user GregorySmith’s findings. I only used the CVS file provided.

Univariate Plots Section

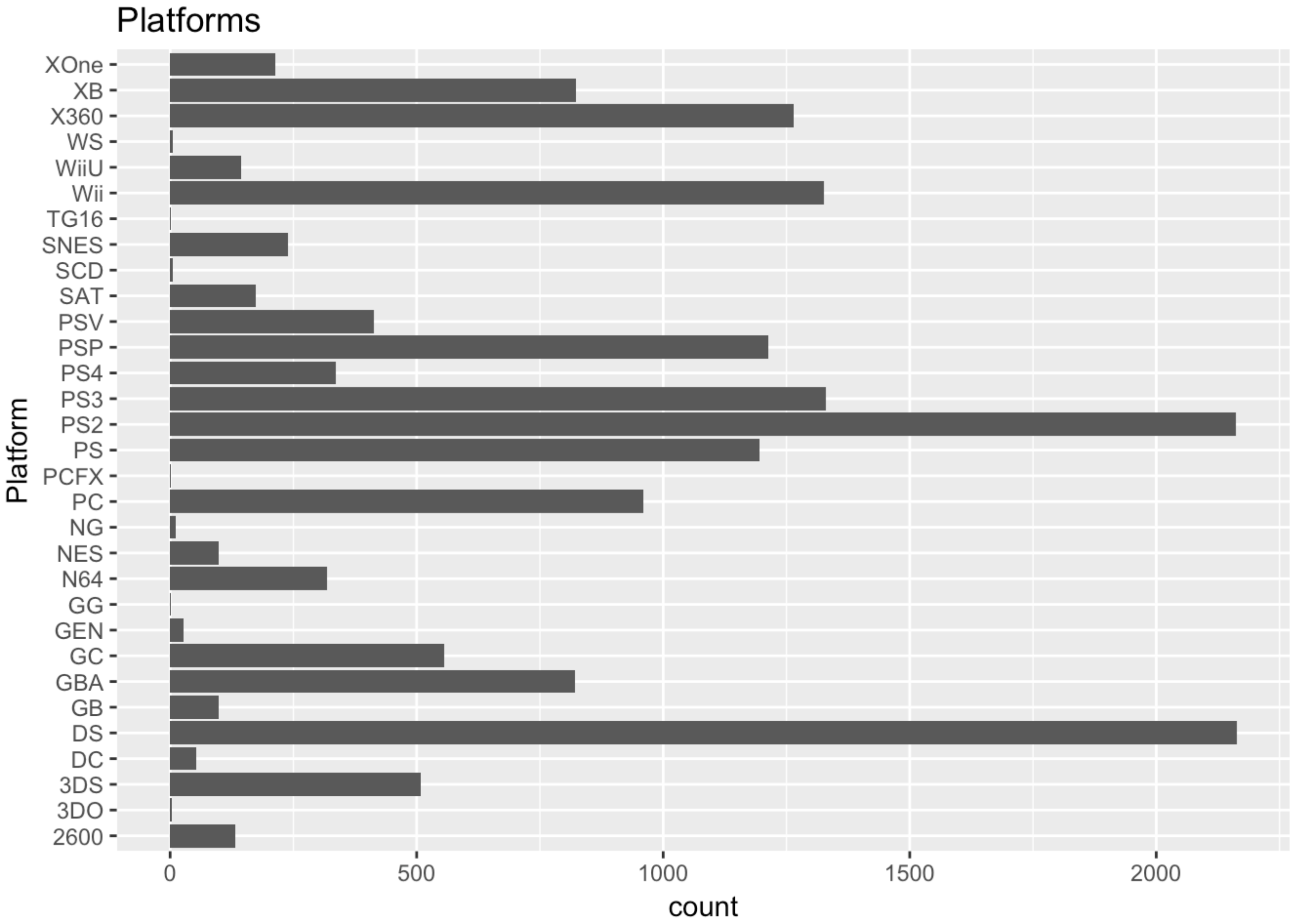
```
##       Rank                                   Name          Platform   
##  Min.   :    1   Need for Speed: Most Wanted:   12   DS     :2163  
##  1st Qu.: 4151   FIFA 14                    :    9   PS2    :2161  
##  Median : 8300   LEGO Marvel Super Heroes   :    9   PS3    :1329  
##  Mean   : 8301   Madden NFL 07              :    9   Wii    :1325  
##  3rd Qu.:12450   Ratatouille                :    9   X360   :1265  
##  Max.   :16600   Angry Birds Star Wars      :    8   PSP    :1213  
##                  (Other)                    :16542   (Other):7142  
##       Year              Genre                             Publisher    
##  2009   :1431   Action      :3316   Electronic Arts             : 1351  
##  2008   :1428   Sports      :2346   Activision                  :  975  
##  2010   :1259   Misc        :1739   Namco Bandai Games          :  932  
##  2007   :1202   Role-Playing:1488   Ubisoft                     :  921  
##  2011   :1139   Shooter     :1310   Konami Digital Entertainment:  832  
##  2006   :1008   Adventure   :1286   THQ                         :  715  
##  (Other):9131   (Other)     :5113   (Other)                     :10872  
##     NA_Sales          EU_Sales          JP_Sales         Other_Sales     
##  Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.00000   Min.   : 0.00000  
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.00000   1st Qu.: 0.00000  
##  Median : 0.0800   Median : 0.0200   Median : 0.00000   Median : 0.01000  
##  Mean   : 0.2647   Mean   : 0.1467   Mean   : 0.07778   Mean   : 0.04806  
##  3rd Qu.: 0.2400   3rd Qu.: 0.1100   3rd Qu.: 0.04000   3rd Qu.: 0.04000  
##  Max.   :41.4900   Max.   :29.0200   Max.   :10.22000   Max.   :10.57000  
##                                                                          
##   Global_Sales    
##  Min.   : 0.0100  
##  1st Qu.: 0.0600  
##  Median : 0.1700  
##  Mean   : 0.5374  
##  3rd Qu.: 0.4700  
##  Max.   :82.7400  
## 
```

```
## 'data.frame':    16598 obs. of  11 variables:
##  $ Rank        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Name        : Factor w/ 11493 levels "¡Shin Chan Flipa en colores!",..: 10991 9343 5531 10993 7364 9707 664
8 10989 6651 2594 ...
##  $ Platform    : Factor w/ 31 levels "2600","3DO","3DS",..: 26 12 26 26 6 6 5 26 26 12 ...
##  $ Year        : Factor w/ 40 levels "1980","1981",..: 27 6 29 30 17 10 27 27 30 5 ...
##  $ Genre       : Factor w/ 12 levels "Action","Adventure",..: 11 5 7 11 8 6 5 4 5 9 ...
##  $ Publisher   : Factor w/ 579 levels "10TACLE Studios",..: 369 369 369 369 369 369 369 369 369 369 ...
##  $ NA_Sales    : num  41.5 29.1 15.8 15.8 11.3 ...
##  $ EU_Sales    : num  29.02 3.58 12.88 11.01 8.89 ...
##  $ JP_Sales    : num  3.77 6.81 3.79 3.28 10.22 ...
##  $ Other_Sales : num  8.46 0.77 3.31 2.96 1 0.58 2.9 2.85 2.26 0.47 ...
##  $ Global_Sales: num  82.7 40.2 35.8 33 31.4 ...
```

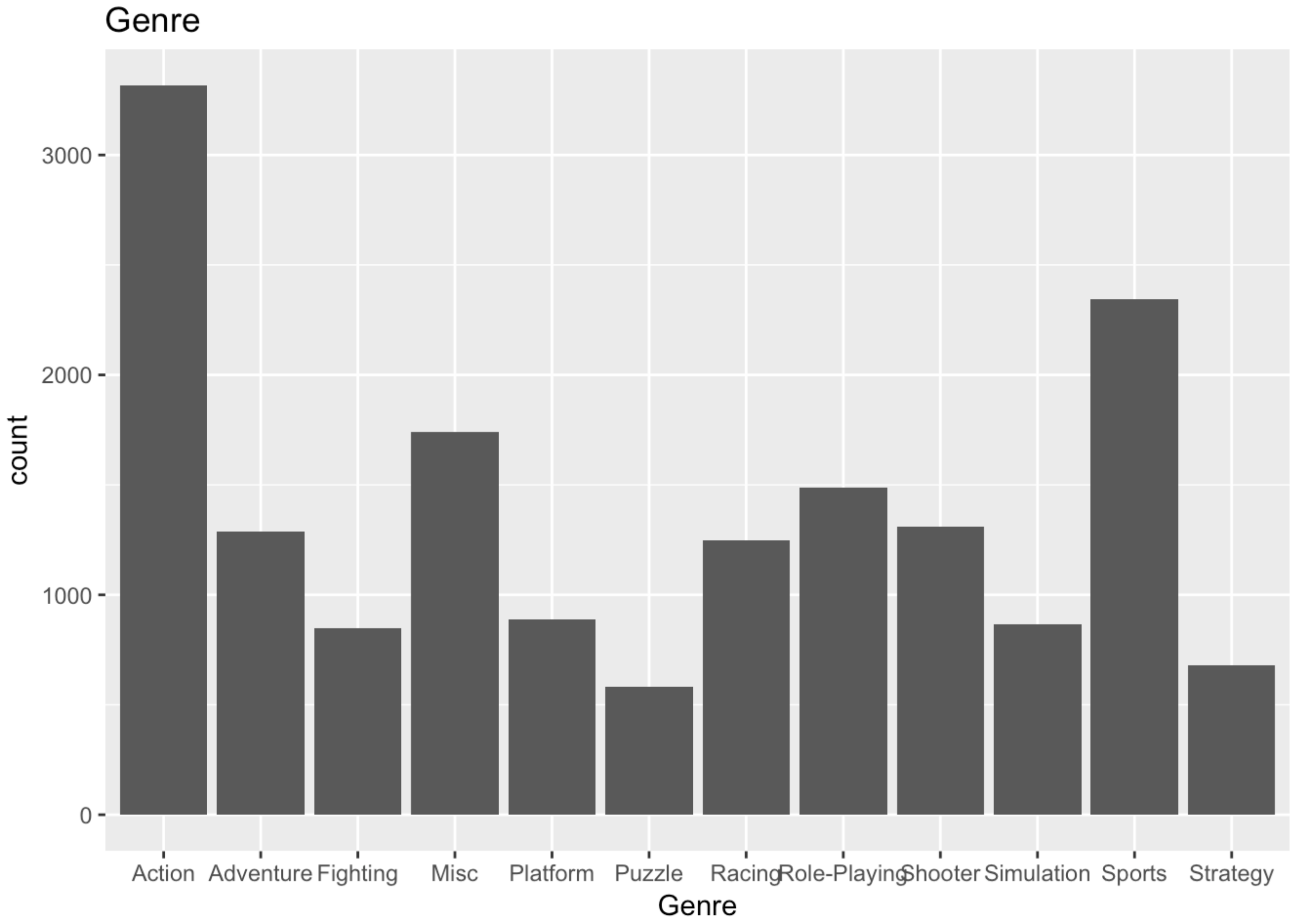
I first started with a summary of the data to get a general understanding of it. We can see that most of the data is comprised of factors, numbers, and characters.



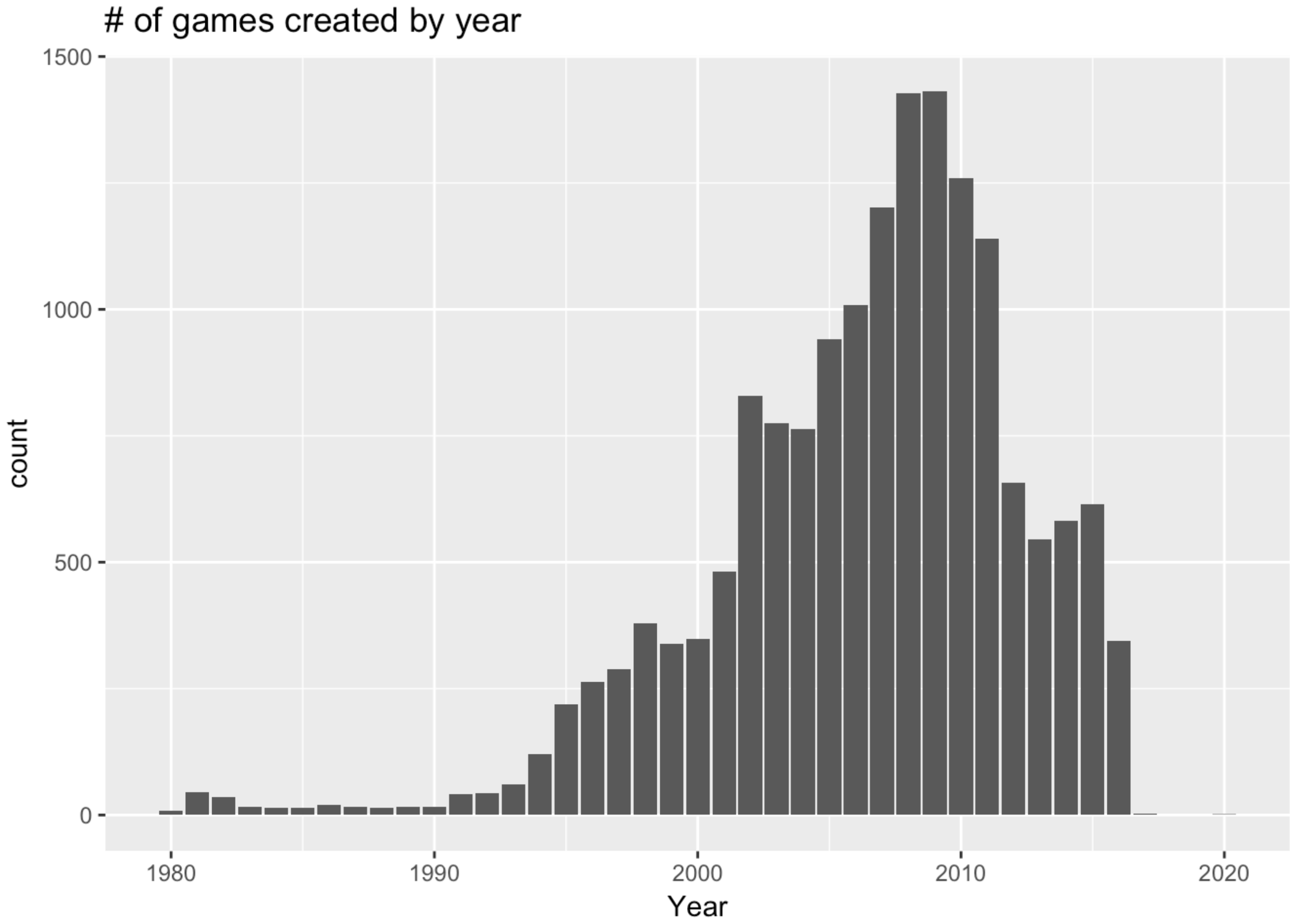
This plot displays how many games publishers have created. After adjusting the bins and limits, we can see that a majority of publishers do not release more then 25 games.



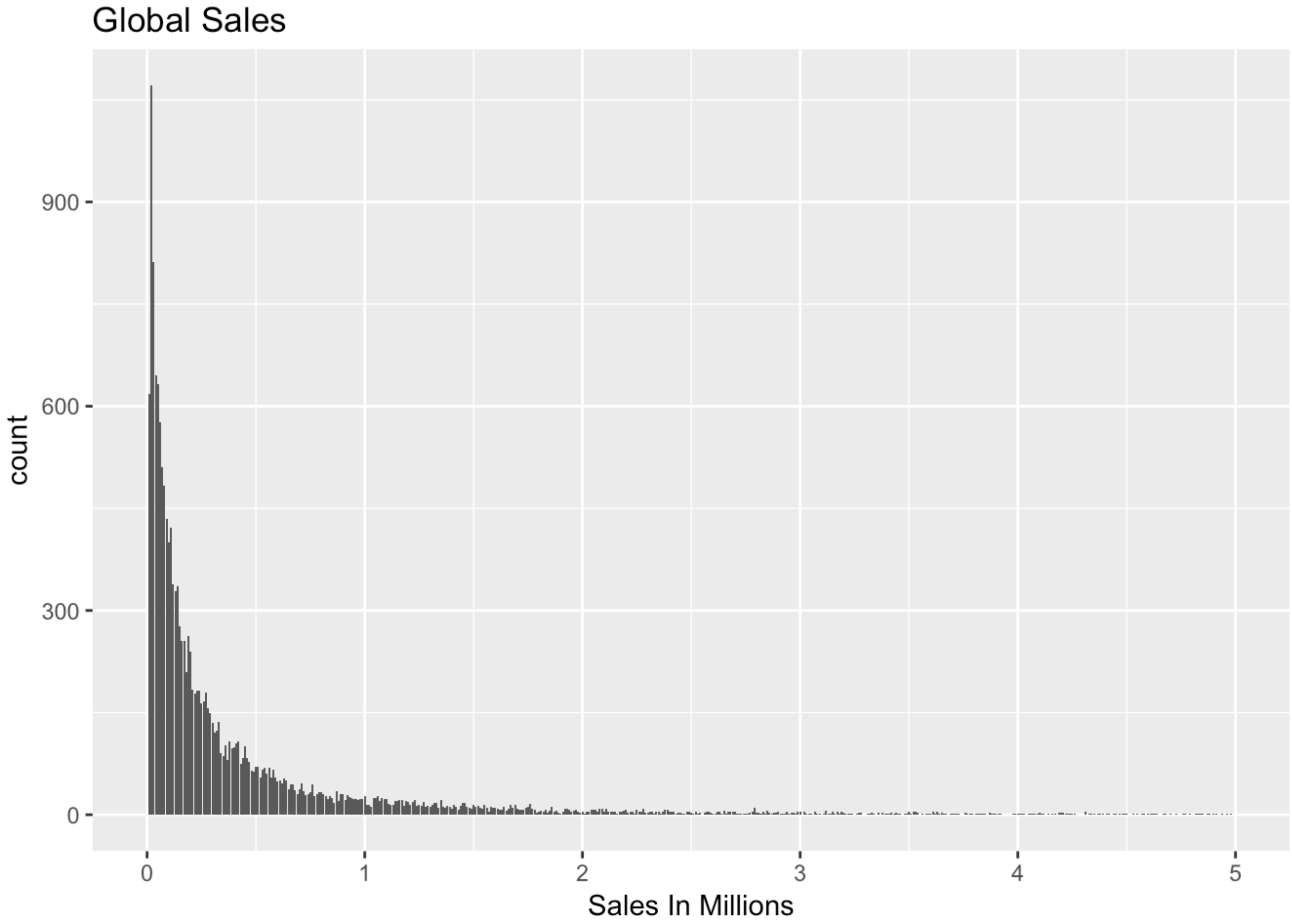
With so many publishers producing different games for different platforms, I wanted to see which platforms did publishers create games for. To my suprise, the DS platfrom had an large amount of games published, considering that the DS has weaker hardware and it’s a hand held platform.



After looking at popular consoles, it’s time to look at which genres are popular. Action and Sports tops the charts. I thought the genres Platform and Puzzle to be higher because these type of genres are older and more easily produced.



This plot shows how many games were release throughout the years. From 2000 to 2010, more games have been release in that decade then the previews 20 years (1980 - 2000) combined. That’s a lot of games.



This plot gives us a general visualizual of how many units a game sells. Most games sell under .5 million units, and it’s exteremely rare for games to sell over 1 million.

Univariate Analysis

What is the structure of your dataset?

There are 16598 observations with 11 variables. Each element is for one game for one platform, which contains:

Rank, (game) Name, Platform, Year (released), Publisher, and 5 sales data: Japan, Europe, North America, Other, Global.

For years, there are 271 elements where the year value is NA.

Some General Observation: - Most publishers will release less than 100 games. - Publisher release games mostly for the PS2 and DS platforms. - Action games are made the most. - A majority of games were release in the 2000 decade. - A majority of sales are undr one million.

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

The main feature I am interested is the sales data. I would like to take a look at how the other variables play a role in sales. Hopefully, examine real world events and try to correlate the data with it.

What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

Platform, Publisher, Genre, Year are the features I would be using when looking at the sales of video games.

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

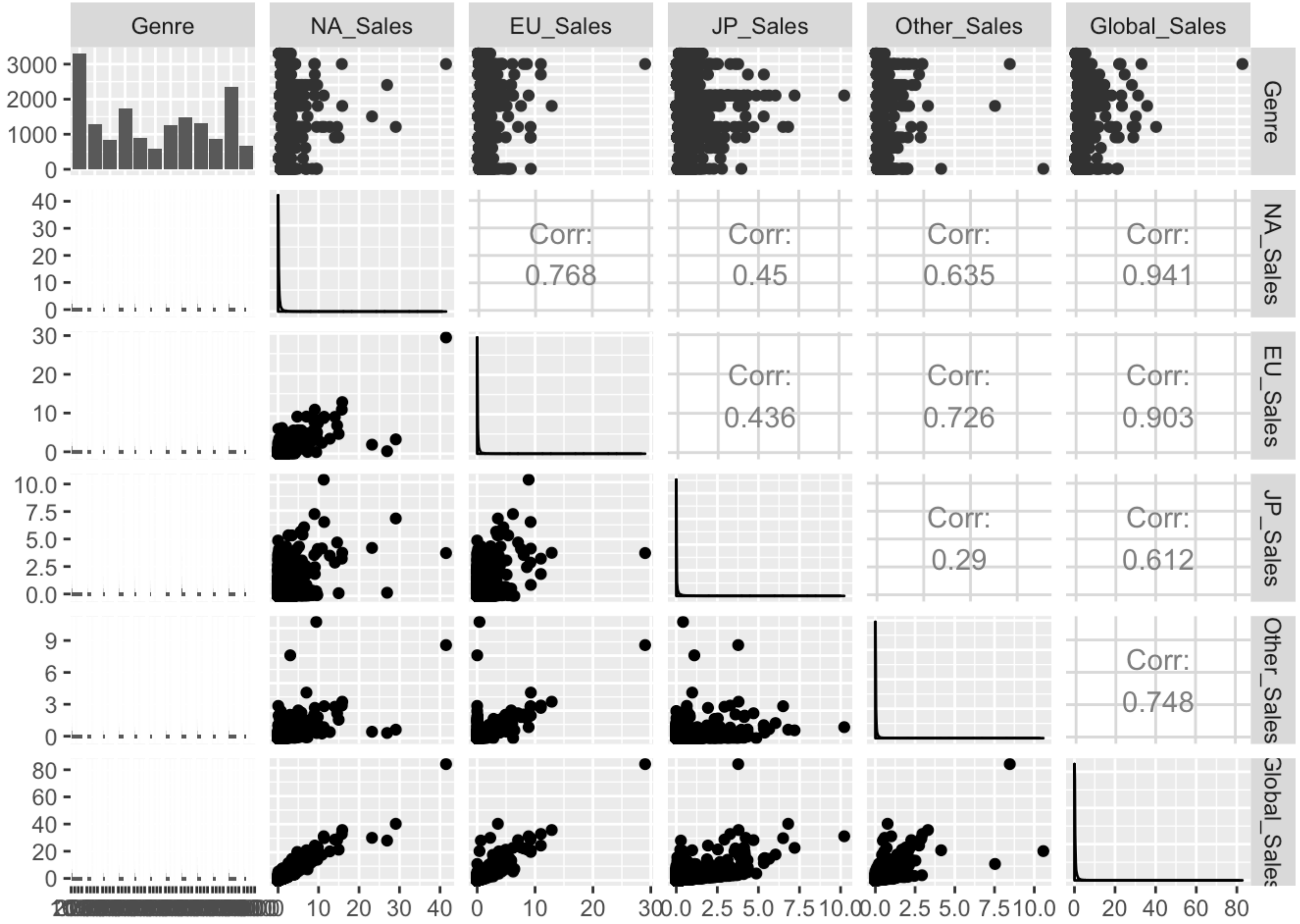
At this point, there was no need to create any new variables to the data. Later on, I will be subsetting and reshap the data when exploring bivariate and multivariate plots.

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

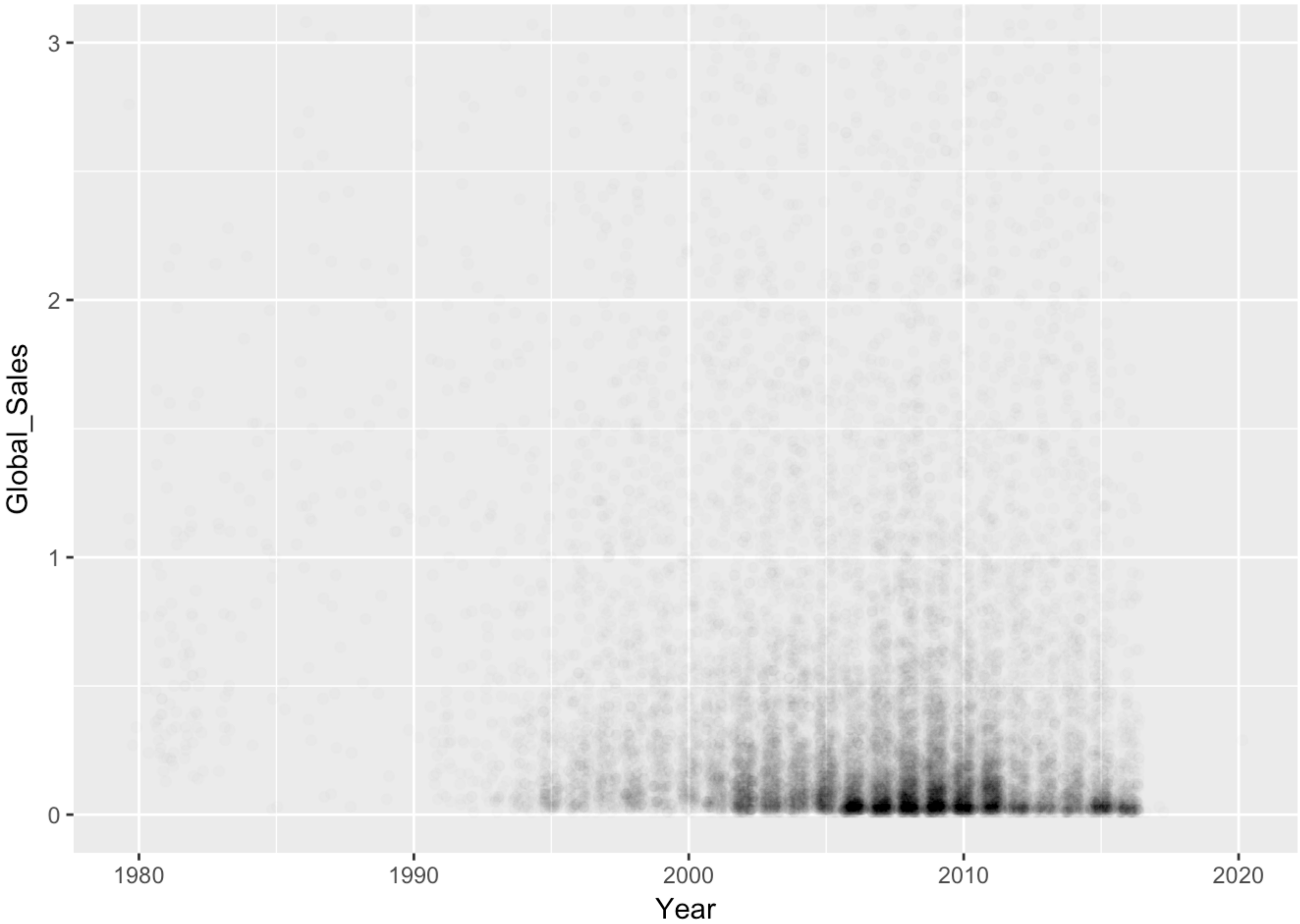
Years is a listed as a factor data type. I changed it to an integer type because it suits my plots better.

The regions could be a little off. Since each region has it’s own population, it could impact how much sales are made.

Bivariate Plots Section

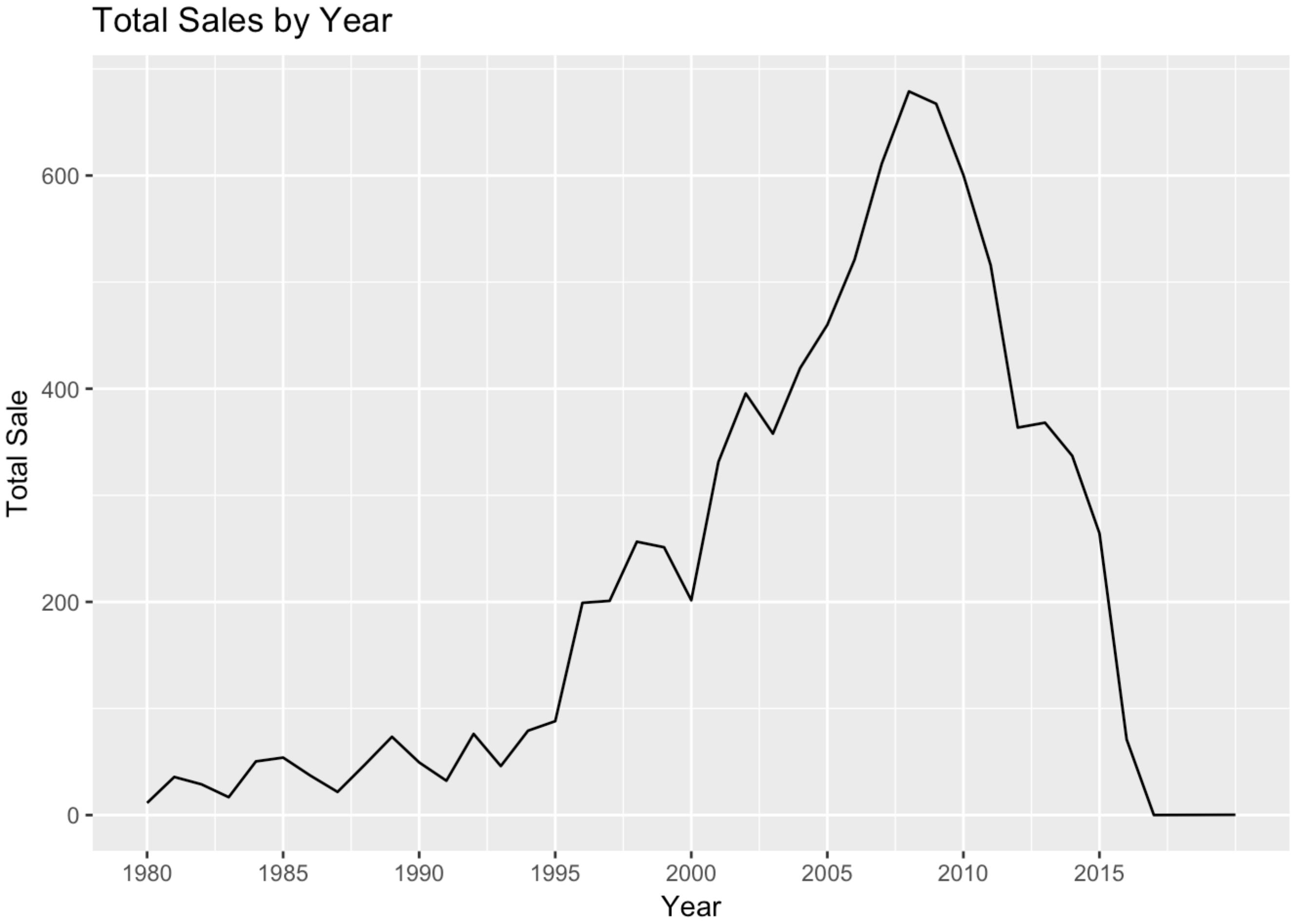


The first thing I noticed with the plot matrix is correlations between global sales and regions. The North American region has the highest while the Japanese region has the lowest, making North America having the highest impact on global sales. Then I looked at the genre and sales section. It is a good indicator that some genre are more popular than others, and each region has different genre popularities.



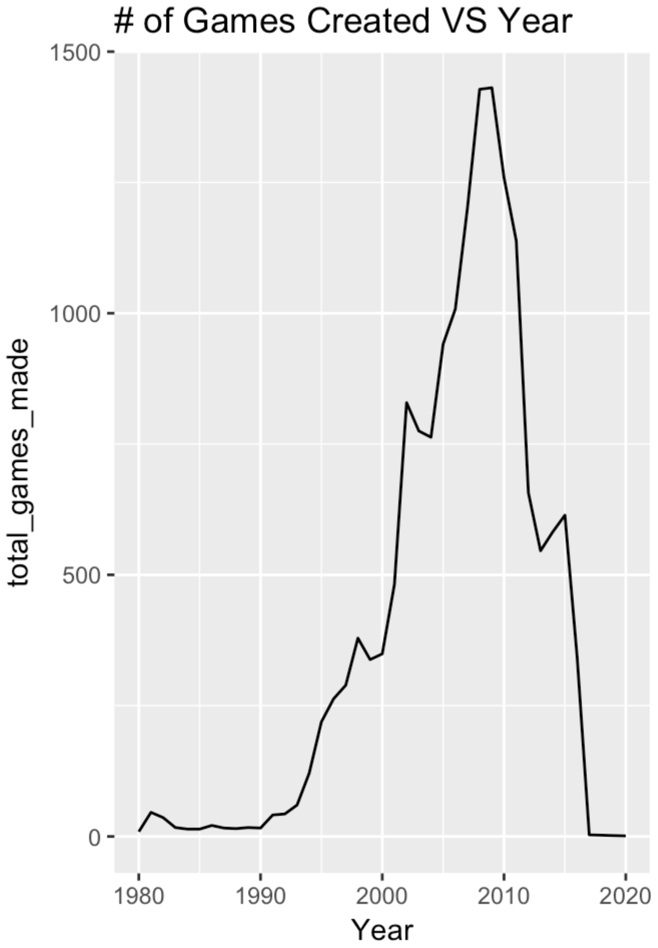
This plot tells a good story about the history of video games. First introduced around the 80’s we see that there are some clusters of sales. It then dies down for about seven years and picks up again around 1995. This could be because of the release of new consoles. The Play Station One (and other consoles) was released around late 1994. We can see a spike in sales in the following years as video games became more popular, until it starts to die down in 2013.

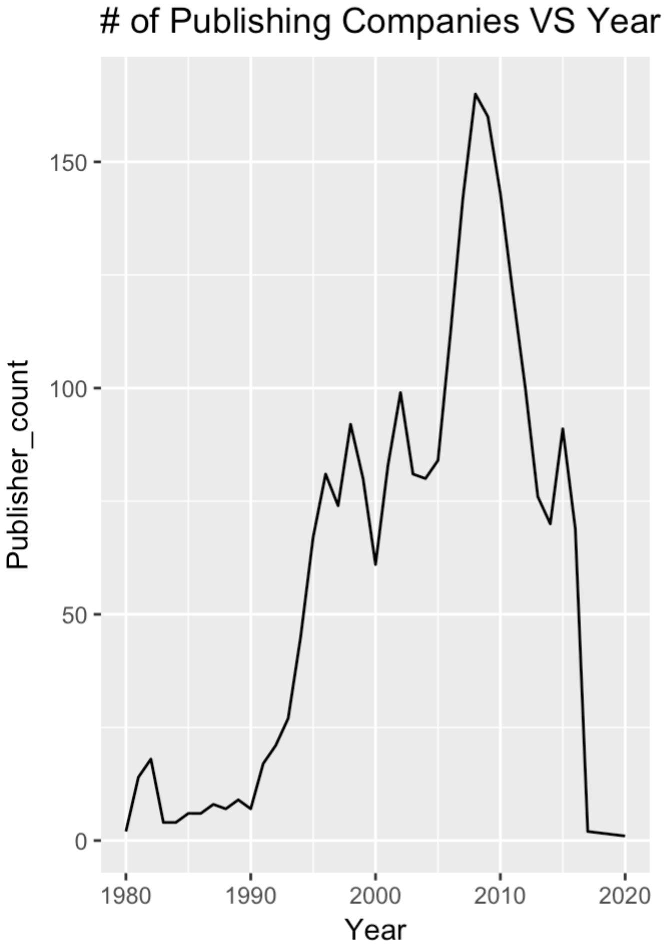
I then went on and created a new data frame that groups the data by year, and summarizes the sum of global sales, sum of publishers, and sum of games created for that year.



Using this data frame, there is a clearer visual of Global Sales. Sales slowly increased since the 80’s until 1995, where it just spikes drastically, tripling in sales. Causes could be because of better technology. As consoles such as the PS1, PS2, PS3, and other generations of consoles are released, there is a spike in the plot.

Sales declined after 2005, and it crashed hard. Even with new consoles could not halt the decline. I decided to take a look at publishing companies and games created to see if those data tells the same story.


# of Games Created VS Year


# of Publishing Companies VS Year

As the crash happens, there are less and less publishing companies, and with less companies, fewer games. A theory for the decline could be because of the great recession which started in 2007 and ended in 2009.

I then took a correlations test, which confirmed that all three categories (sales, publishers, games created) are correlated with each other. So when one declines, so does the others.

```
# the plot and cor.test shows that there is a direct correlation between numbers of games created and sales.

with(vg_data, cor(sales.by.year$sum_sale, sales.by.year$total_games_made))
```

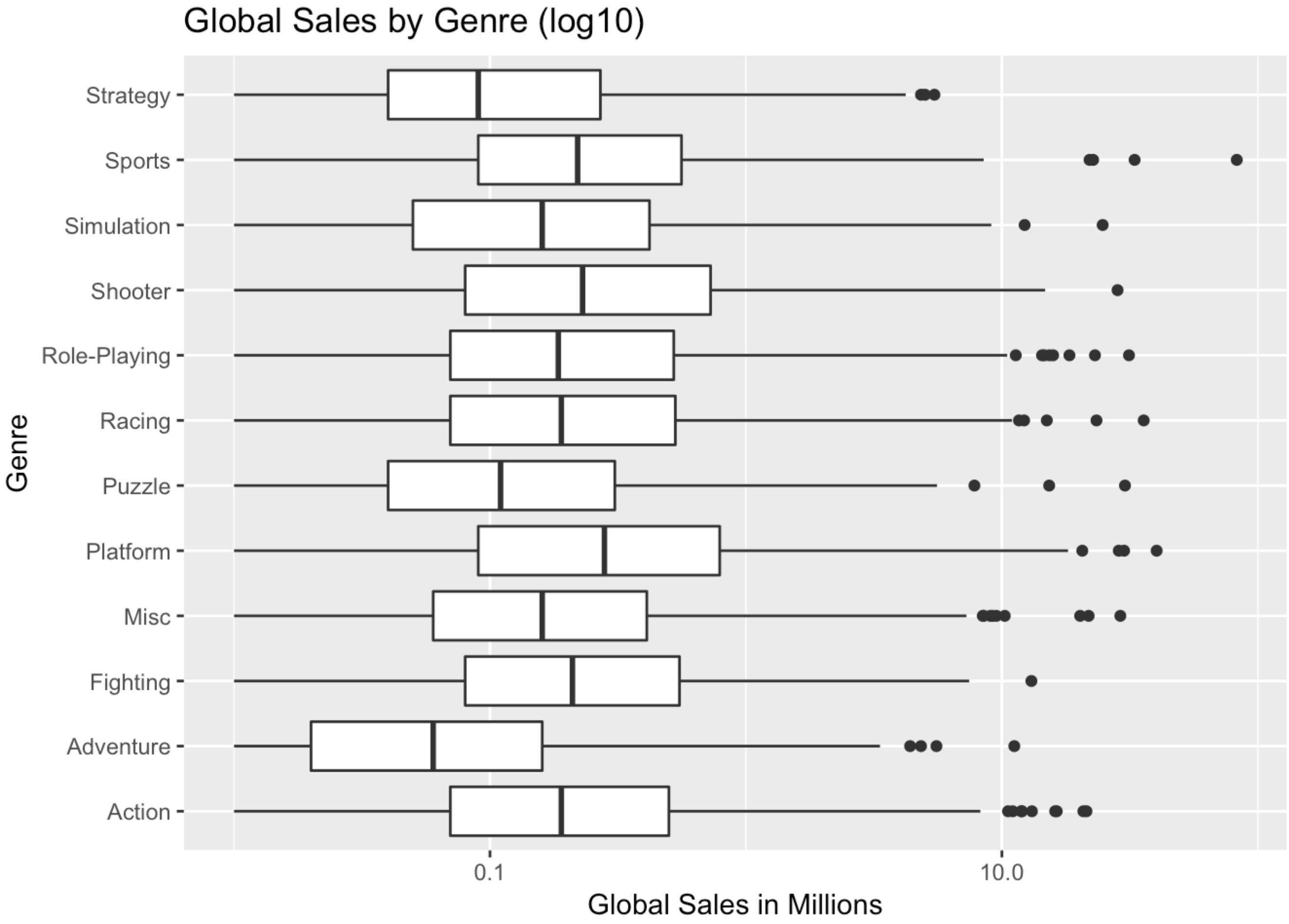
```
## [1] 0.9822252
```

```
with(vg_data, cor(sales.by.year$sum_sale, sales.by.year$Publisher_count))
```

```
## [1] 0.9417013
```

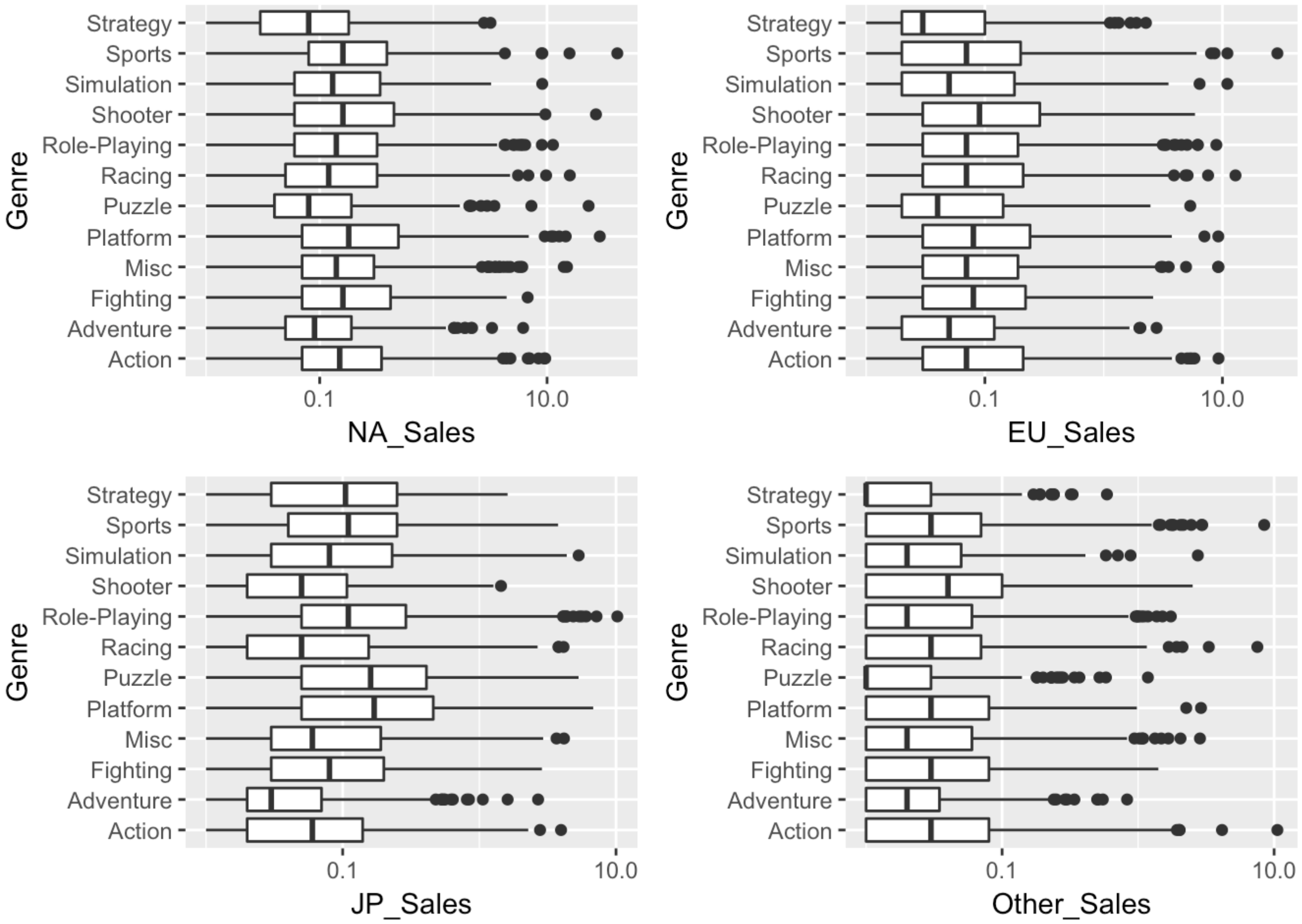
```
with(vg_data, cor(sales.by.year$Publisher_count, sales.by.year$total_games_made))
```

```
## [1] 0.9438532
```



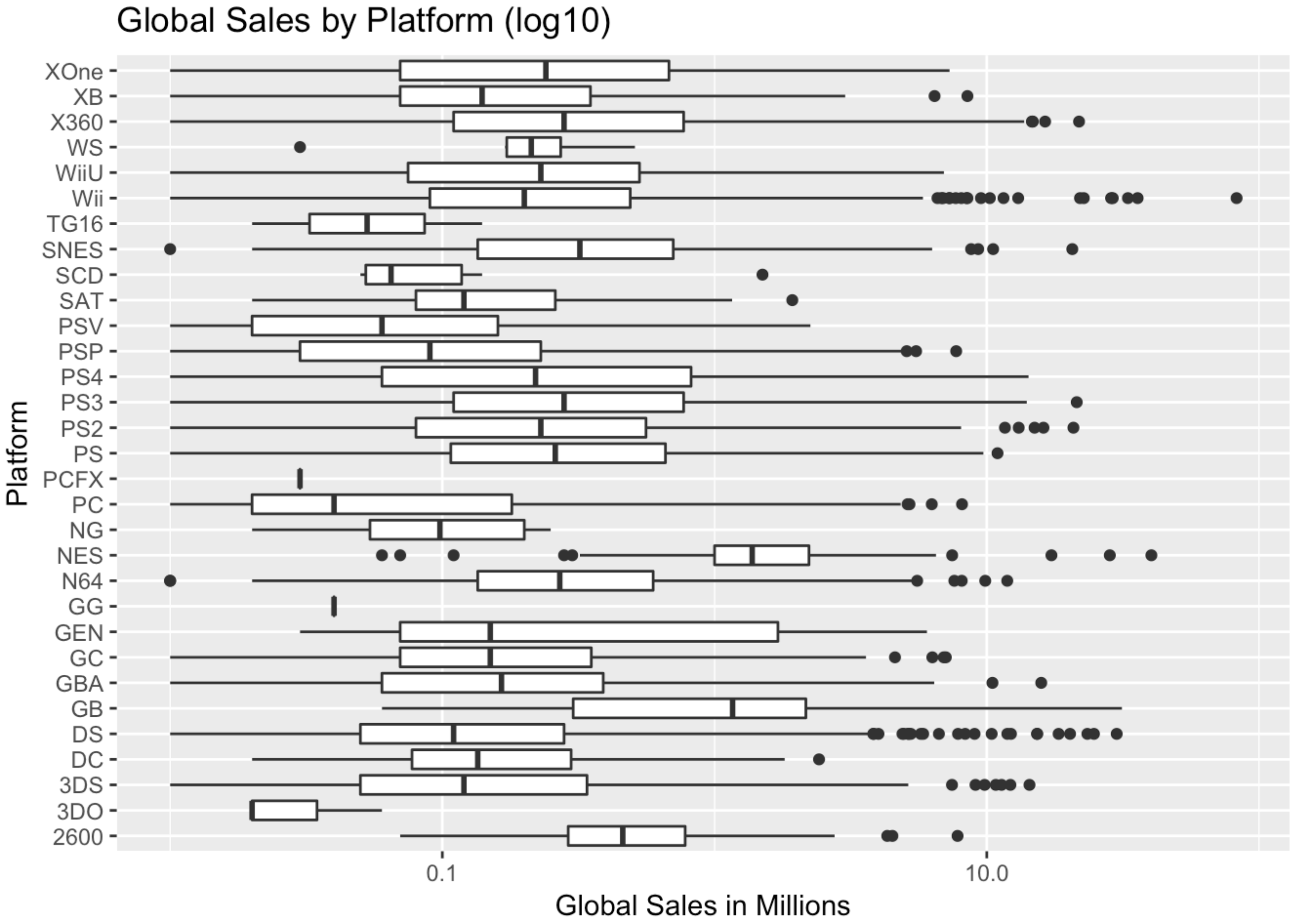
Comparing global sales and genre gives a general idea of which genre are popular.

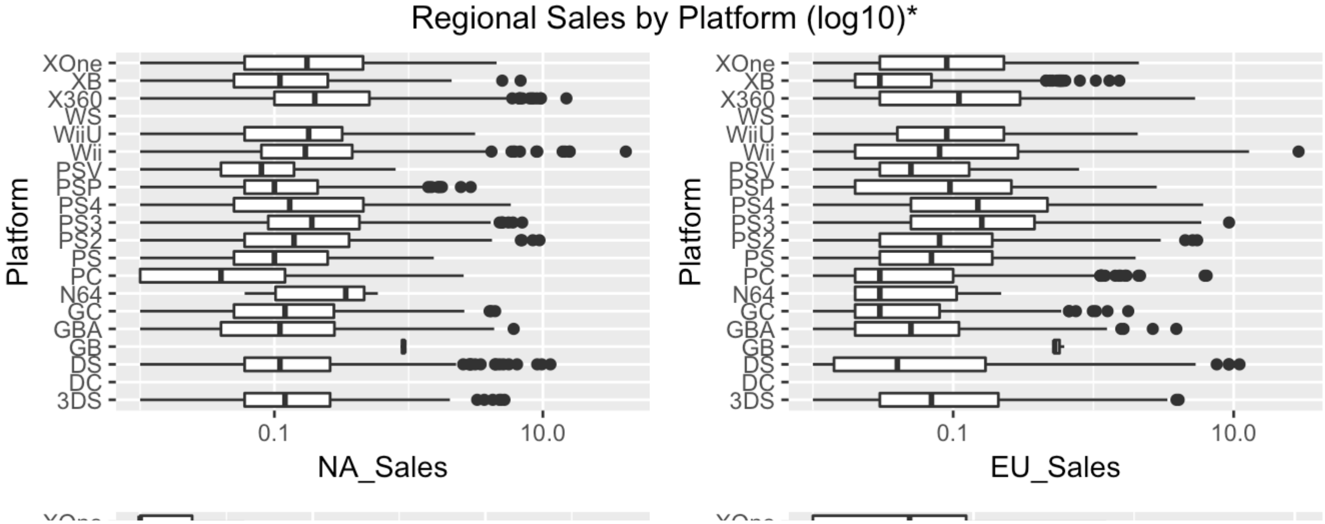
It should be noted that some games can have more than one genre (ie: Resident Evil 5 is both a shooter (third person shooter) and action, but is only labeled as Shooter in the data). We can assume that if the popular a genre is, the more sales it will have, and we can see that Shooters, Action, and Role-Playing games are pretty popular globally.

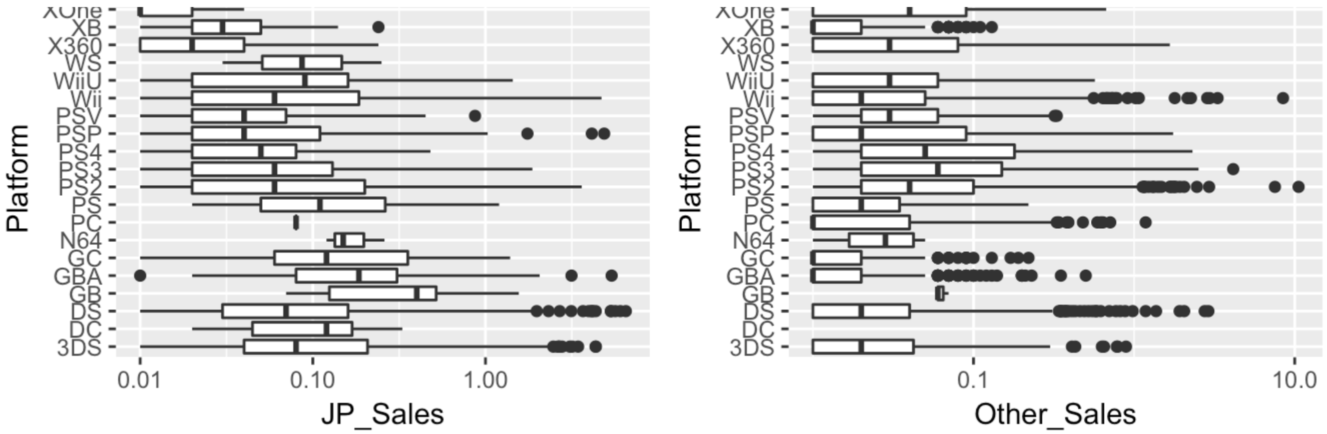


When looking at genre and sales by region, we can see different popularity in games. Platform games are very popular in Japan, while North America prefers shooters and platform games.

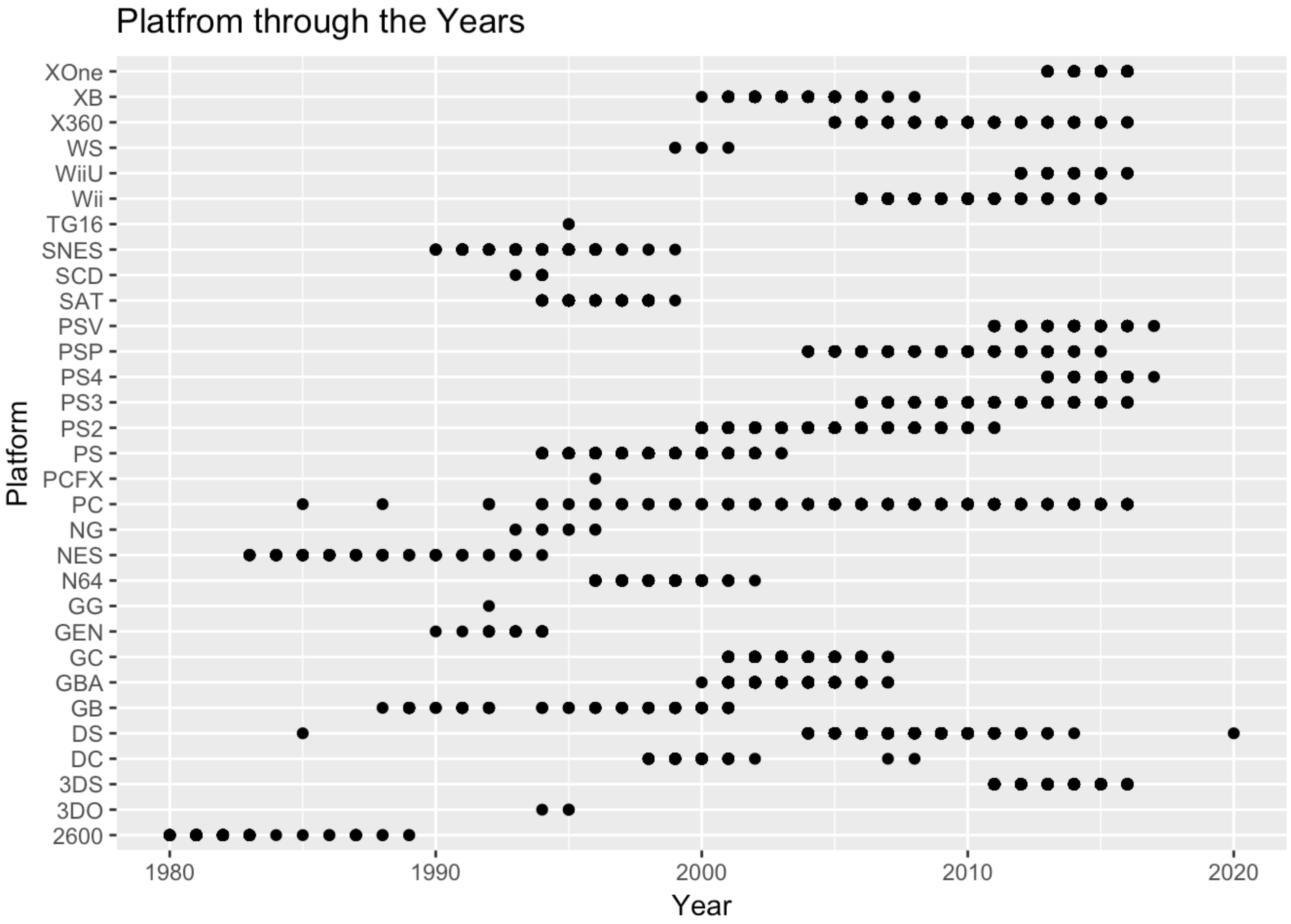
I created series of similar plots below showing which platforms are popular globally and regionally.


Global Sales by Platform (log10)
Platform
XOne
XB
X360
WS
WiiU
Wii
TG16
SNES
SCD
SAT
PSV
PSP
PS4
PS3
PS2
PS
PCFX
PC
NG
NES
N64
GG
GEN
GC
GBA
GB
DS
DC
3DS
3DO
2600
0.1
10.0
Global Sales in Millions


Regional Sales by Platform (log10)*
Platform
XOne
XB
X360
WS
WiiU
Wii
PSV
PSP
PS4
PS3
PS2
PS
PC
N64
GC
GBA
GB
DS
DC
3DS
0.1
10.0
NA_Sales
Platform
XOne
XB
X360
WS
WiiU
Wii
PSV
PSP
PS4
PS3
PS2
PS
PC
N64
GC
GBA
GB
DS
DC
3DS
0.1
10.0
EU_Sales
XOne
XOne



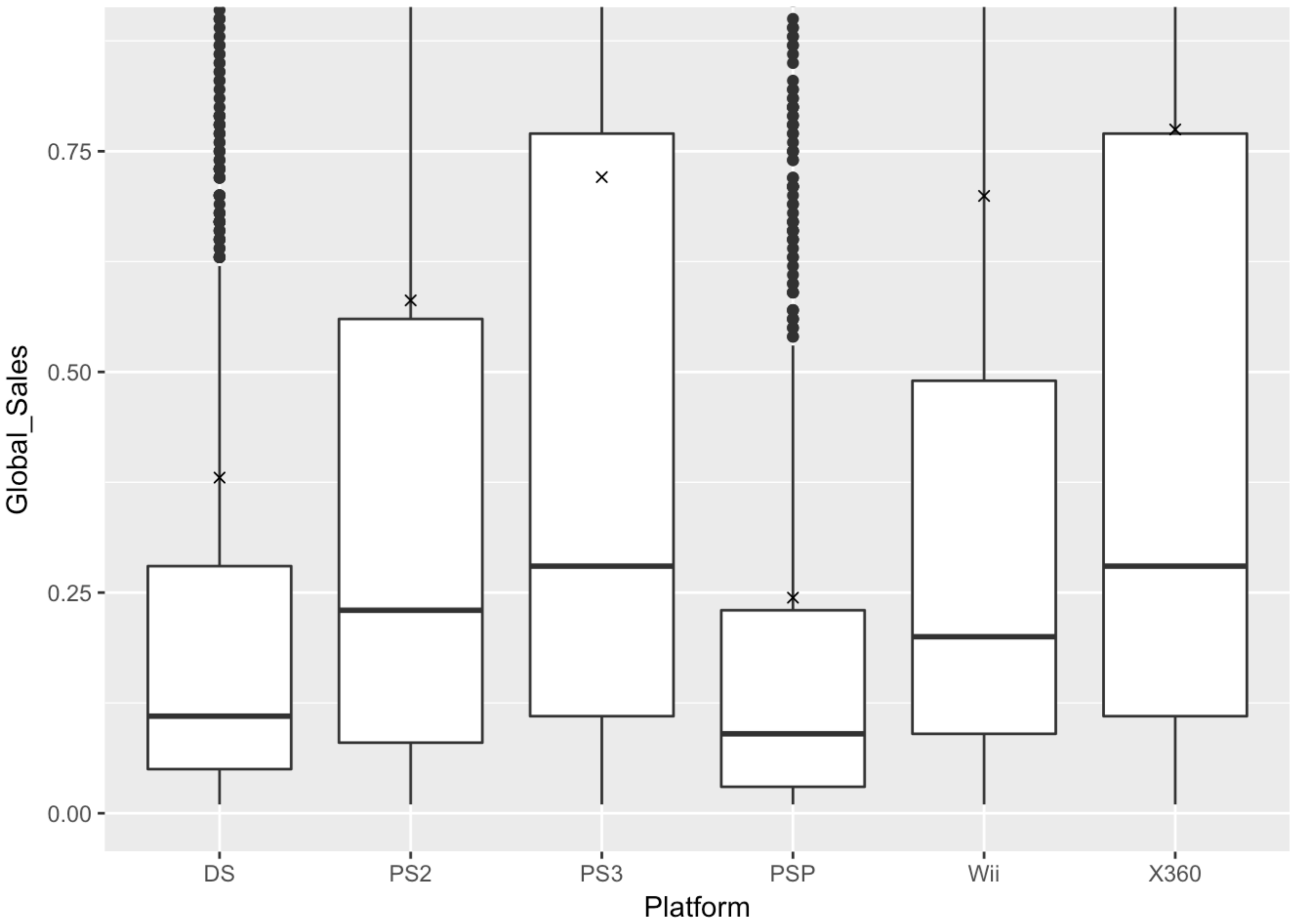
*There are too many platforms to plot. Data shown are platforms created in 2000 and newer



This scatters plot gives a good visual of the life span for platforms, with each point representing a year. We can also see a discrepancy with the DS platform, having a game in 1985 and 2020.

Next, I wanted to look at the top 6 platform with the most amount of games. Below is a chart displaying the total amount of games for a platform.

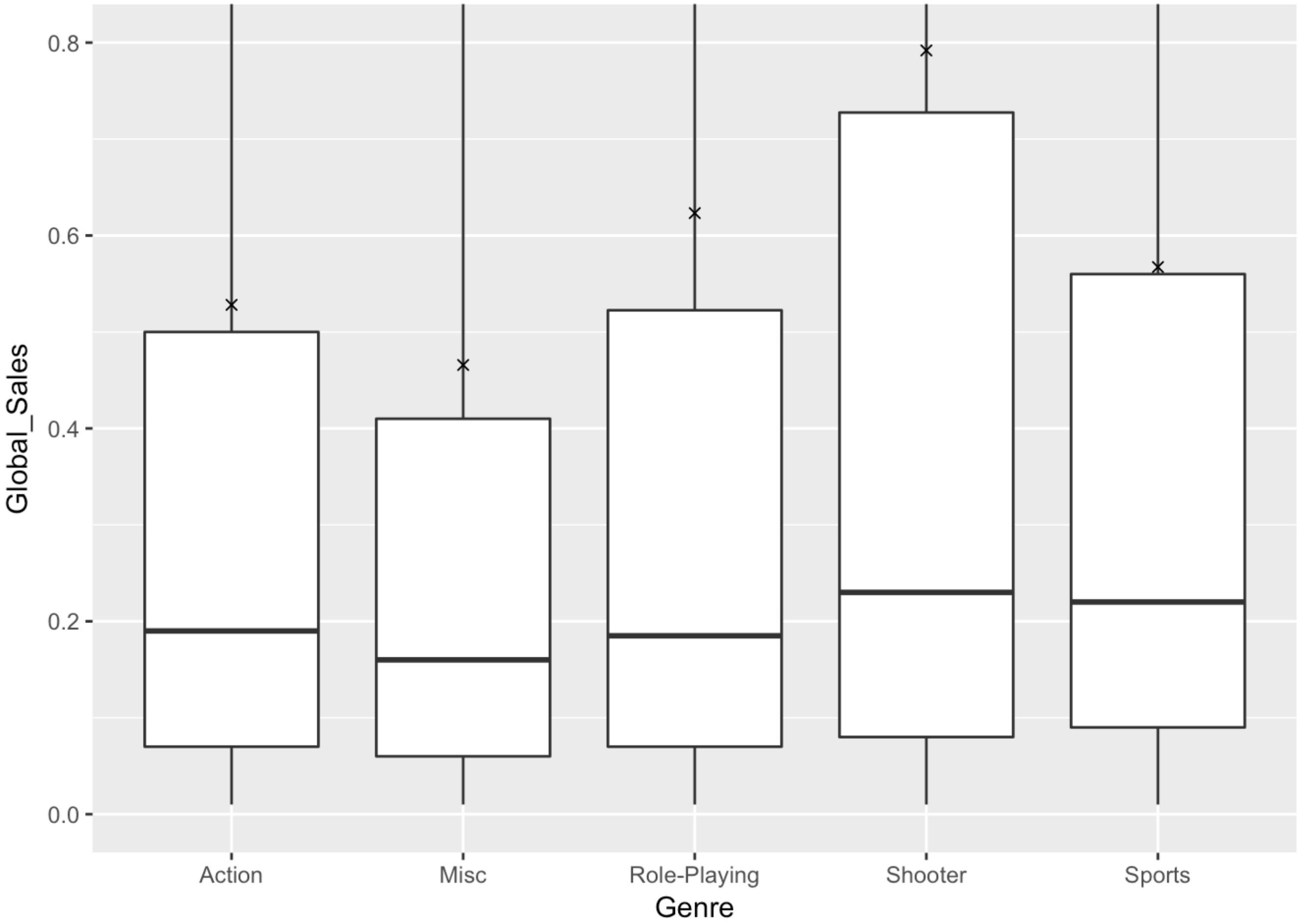
```
## 
## 2600  3DO  3DS   DC   DS   GB  GBA   GC  GEN   GG  N64  NES   NG   PC PCFX 
##  133    3  509   52 2163   98  822  556   27    1  319   98   12  960    1 
##   PS  PS2  PS3  PS4  PSP  PSV  SAT  SCD SNES TG16  Wii WiiU   WS X360   XB 
## 1196 2161 1329  336 1213  413  173    6  239    2 1325  143    6 1265  824 
## XOne 
##  213
```



Here we can see that a majority of games for the top six only sells under one million units. What surprised me most is that the mean (show by the ‘x’ mark) are in the 75% quantile and above for all six consoles. Another interesting note is that the PS and DS have thinner boxes, and fewer sales when compared to others. This is a small preview of how handheld consoles fair VS more powerful console systems.

I did a similar plot about the top six genres.

```
## 
##       Action    Adventure     Fighting         Misc     Platform 
##         3316         1286          848         1739          886 
##       Puzzle       Racing Role-Playing      Shooter   Simulation 
##          582         1249         1488         1310          867 
##       Sports     Strategy 
##         2346          681
```



Mosts genres did not sell more than .6 million units, with the median being in 200,000 units. When comparing this plot to the platforms box plot, I noticed that there was also less variance in sales as well.

Bivariate Analysis

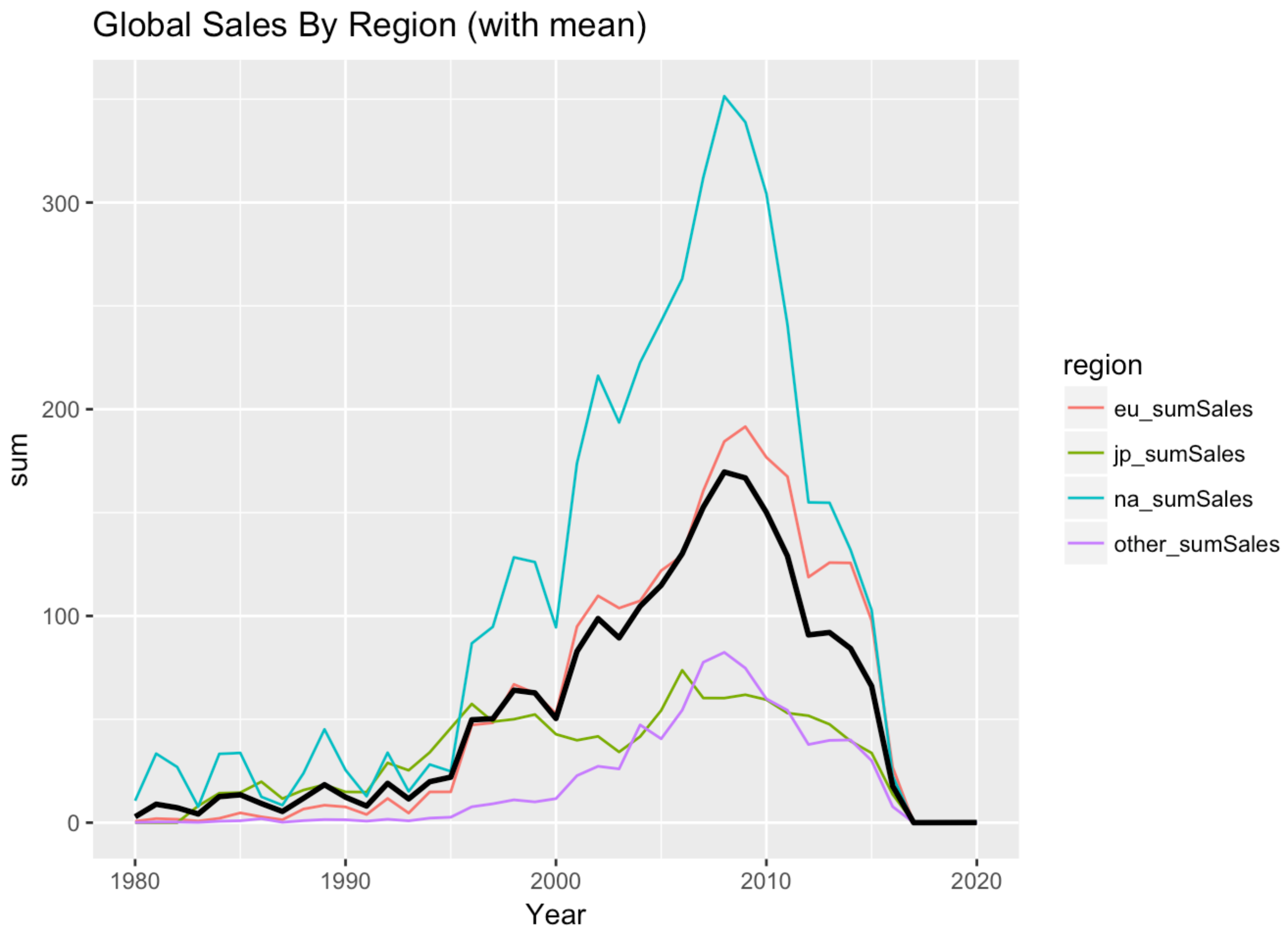
Talk about some of the relationships you observed in this part of the

1. Publishing Companies, number of games created (yearly), and sales are all correlated with each other. With more publishing companies, there can be more games, thus more sales. Perhaps the recession caused a lot of studios to close, impacting games and sales and explaining the decline in 2005.
2. Sales of genre show how games of different genre are popular in different regions. Action and shooters are popular in the North American Region, while Role-Playing, Sports, and Action are popular in the Japanese region.
3. Consoles made by Japan seems to be the dominate console in each region. Although I did find it interesting that PC games were not as high. Since PC's are accessible (and has more uses), one would assume that it would have consistent sales throughout each region.
4. It is uncommon for a game to make over 1 million sales. Most games will have about 200,000 units sold. So when a game reaches over a million units sold, it speaks about how well that game was made.

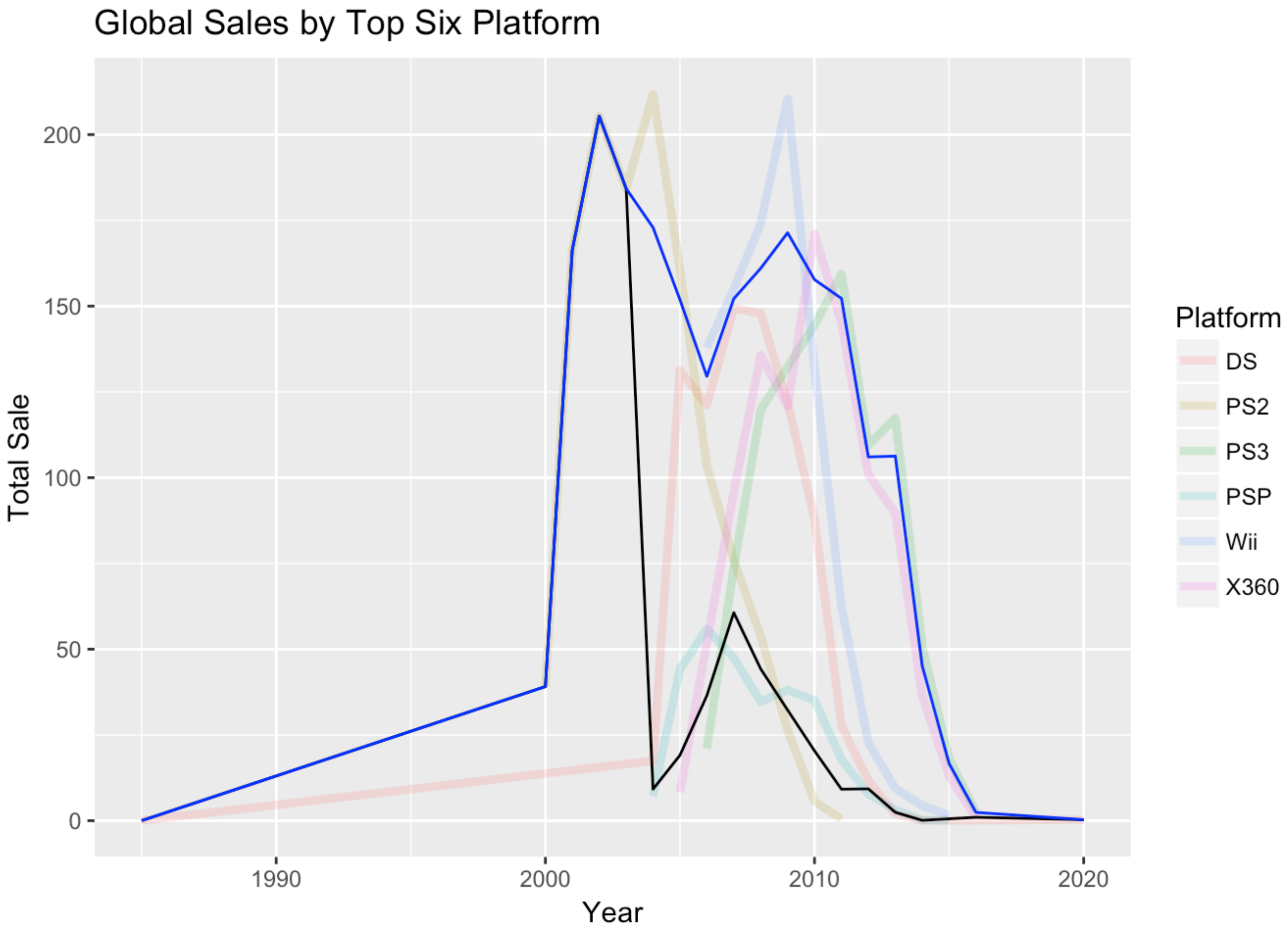
What was the strongest relationship you found?

The number of publishing companies, games created, and sales are strongly correlated, with games created and sales being the strongest. I thought publishing companies and sales would be more strongly correlated because publishing companies are the ones who profit the most.

Multivariate Plots Section



When dividing global sales by region, we can see that North America accounts for the majority of the data. While the region Other and Japan performs below the average. Interestingly enough, Europe performs near the average in sales.

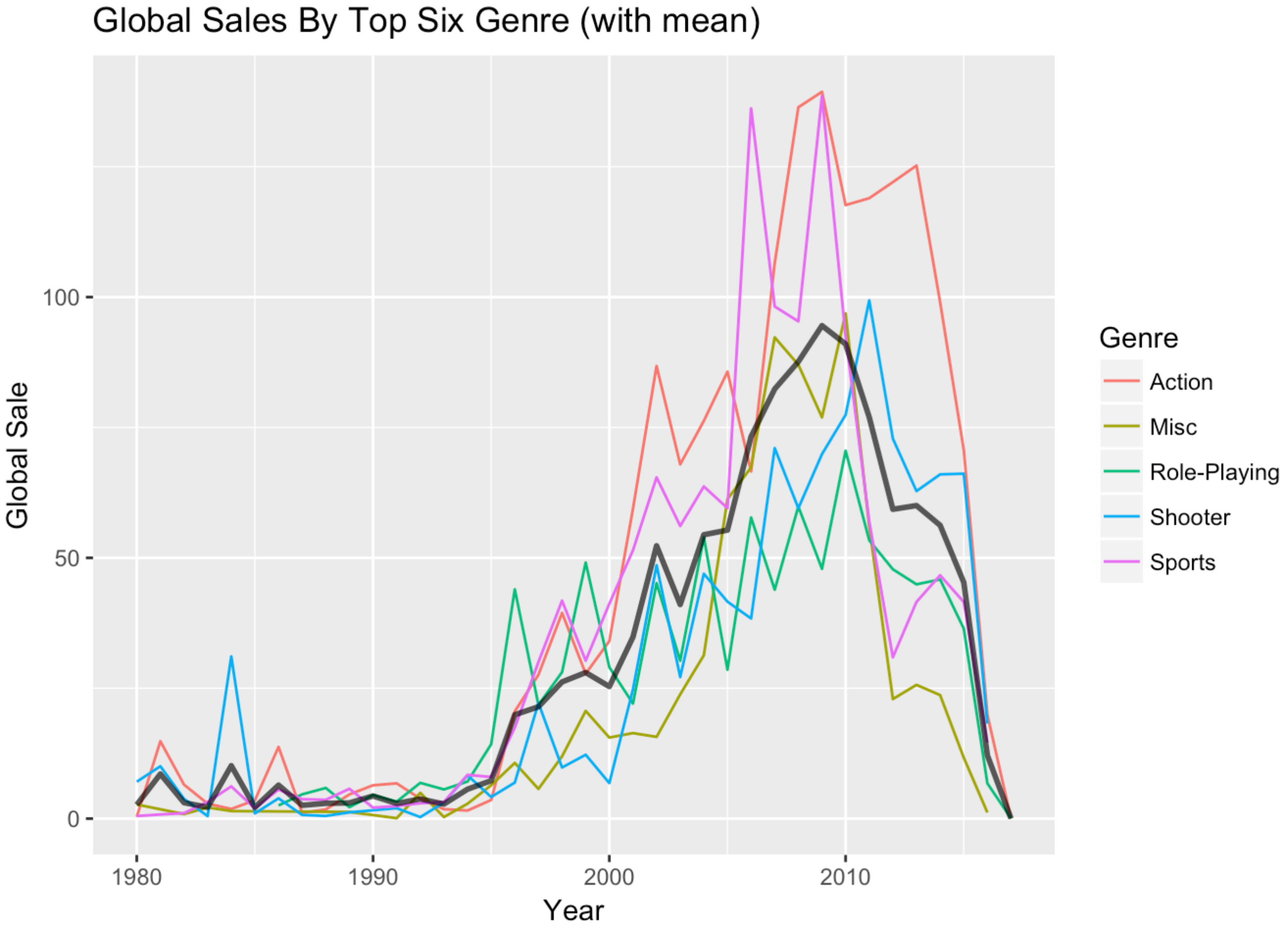


Here we can see sales by consoles, with the blue line signifying the 90th percentile and the black line as the 10th percentile. We can see that only three consoles sold over the 90th percentile at for a certain time over the years.

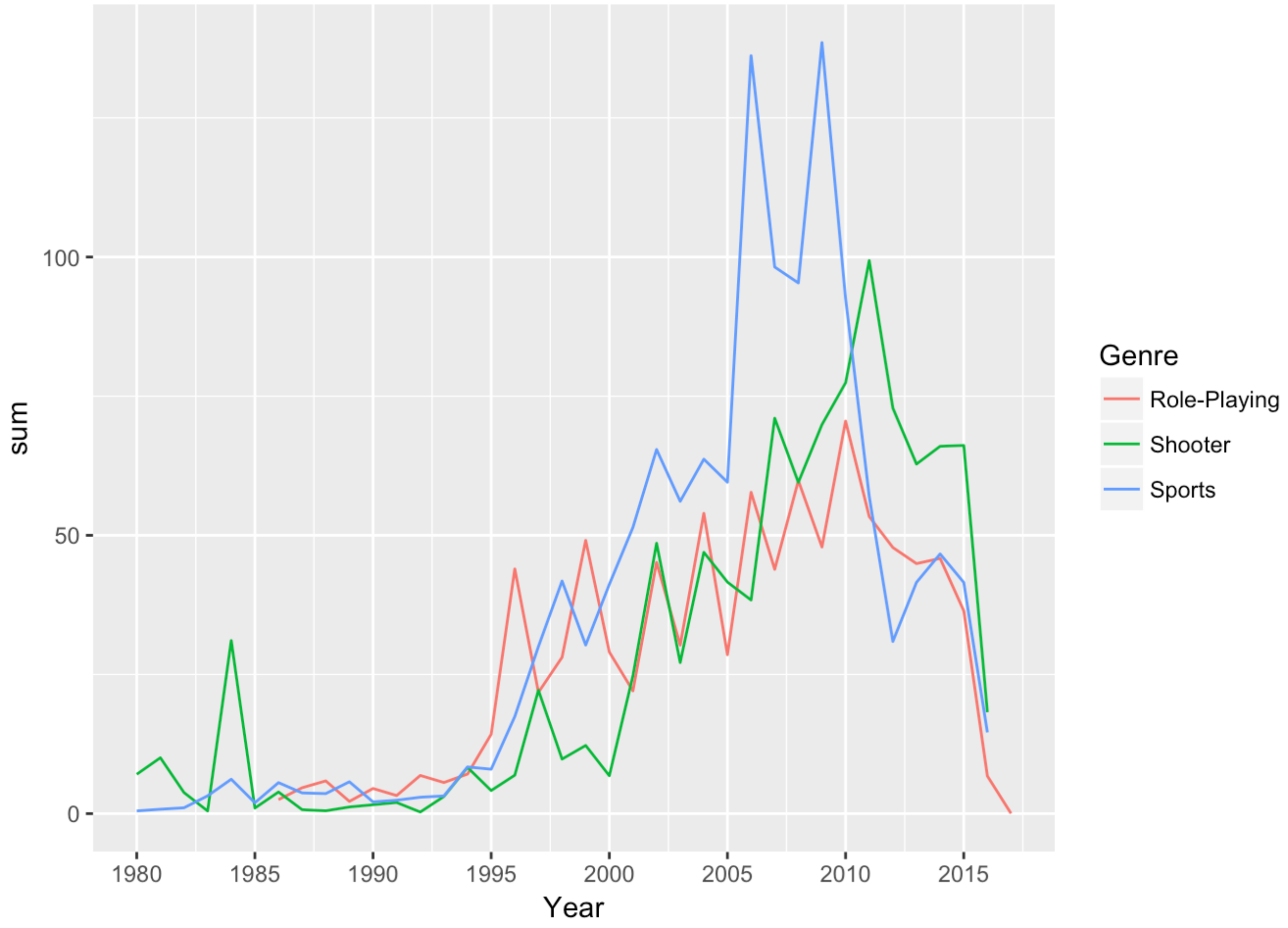
Note: Even though the DS was released in 2004, apparently there was a DS game released in 1986 (according to GameWise.co).

Here are some statistics about the top six consoles.

```
## top_six.console$Platform: DS
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0100  0.0500  0.1100  0.3803  0.2800 30.0100 
## --------------------------------------------------------- 
## top_six.console$Platform: PS2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.010   0.080   0.230   0.581   0.560  20.810 
## --------------------------------------------------------- 
## top_six.console$Platform: PS3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0100  0.1100  0.2800  0.7207  0.7700 21.4000 
## --------------------------------------------------------- 
## top_six.console$Platform: PSP
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0100  0.0300  0.0900  0.2443  0.2300  7.7200 
## --------------------------------------------------------- 
## top_six.console$Platform: Wii
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0100  0.0900  0.2000  0.6994  0.4900 82.7400 
## --------------------------------------------------------- 
## top_six.console$Platform: X360
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0100  0.1100  0.2800  0.7747  0.7700 21.8200
```

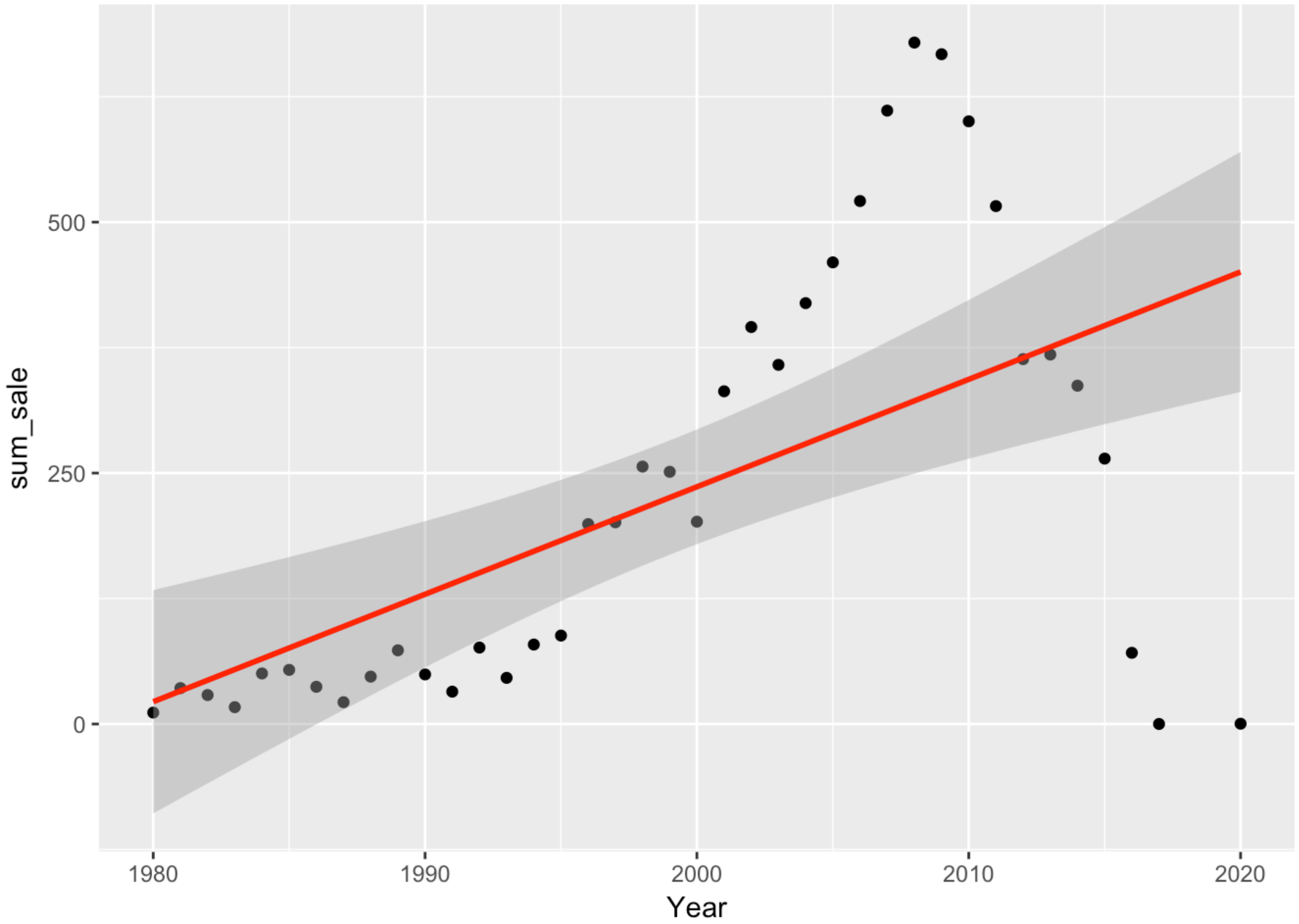


When looking at this plot, we see that Action and Sports were are the top two genre's, performing higher than the average for every year after 1996.

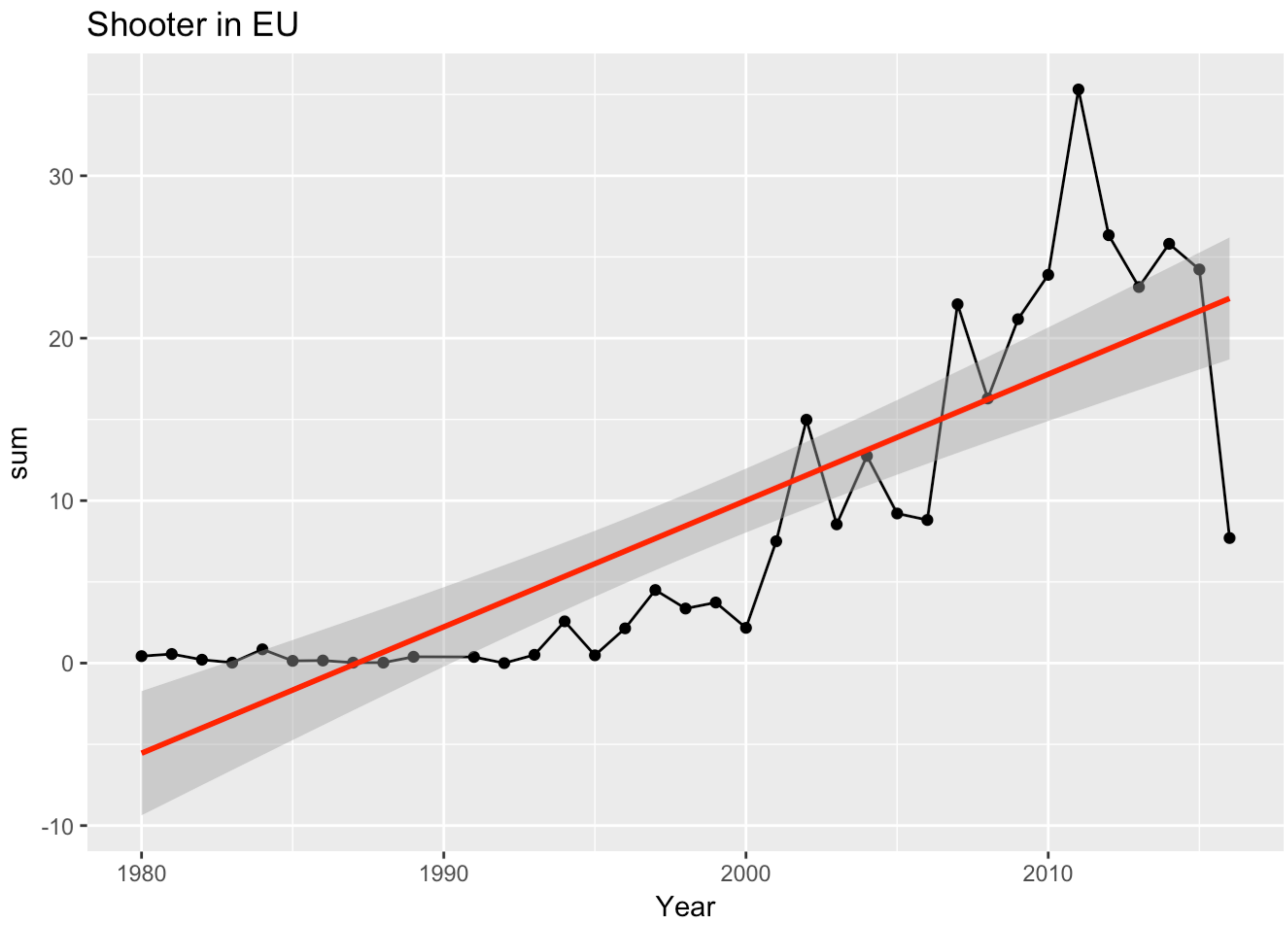


I wanted to take a look at these three genres because they all have spikes in them throughout the years. I looked into it and search for possible reasons for the spikes. Role-Playing and Shooters have spiked because it has years where blockbuster games such as Call of Duty were released, peaks, then decline until the next blockbuster was released.

The spikes for sports could be caused by the FIFA world cup, and during years where the World Cup was taking place, sports games started to increase.



Using a linear regression we can predict sales for the future. Even though there was the sharp decline, our data shows that there is still a positive relationship between sales and years, and the linear model predicts that sales will still increase.



I wanted to demonstrate that a linear model can be used to also predict how well a genre will do for a specific region. In this case, Shooters would perform quite well in the EU.

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?

Much of my data was focusing on the global sales between region, platform, and genre. A noticeable trait with global sales is that the North American region accounts for a majority of games sales. Only European sales remained close to the mean sales, while the other regions were under the mean.

With the top six platforms, I noticed discrepancies with the DS data. The console was created in 2004, and yet there was data for in the 80’s. What surprised me the most was that the Wii platform had higher maximum sales than it rivals, but the PS3 and Xbox 360 oure preformed the Wii.

Taking the top six genres, sports and action had the highest. What interested me the most was exploring possible causes for these spikes. For example, Sports has two major spikes, and during those two spikes, the FIFA world cup was that year. World events have an effect on sales and we take these into account while exploring the data.

Were there any interesting or surprising interactions between features?

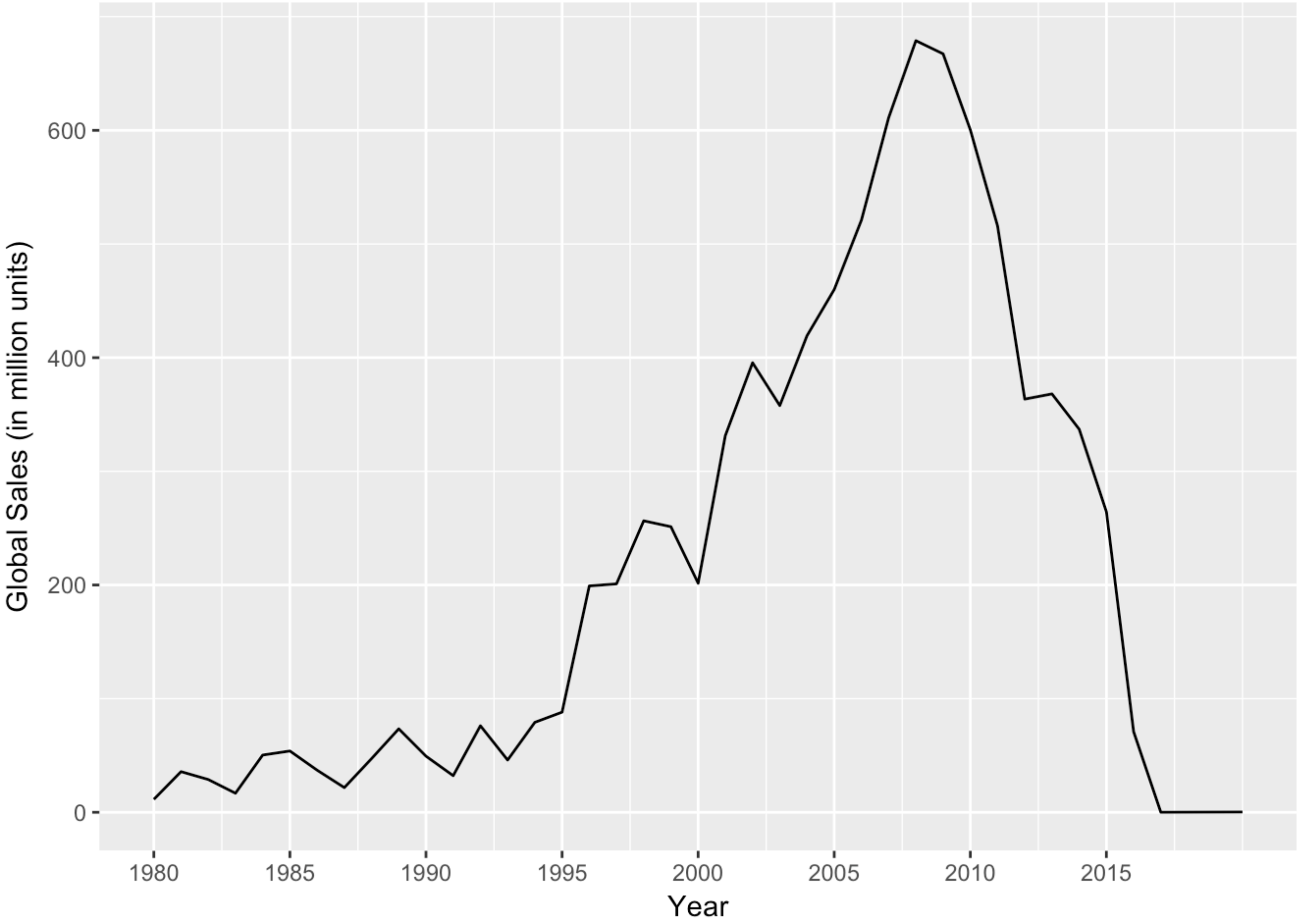
What suprised me the most how the world plays a huge roles with sales. Around 2006, everything starts to rapidly dececline. During this time, the recessison was occuring and this had an affect on everything, regaurdless of region, platform, and/or genre.

OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

I created two models. One to predict how the sales of video games will look in the future, and another one to predict how well a genre would for a specific region. A strength of using a linear model is that we can use it as financial predictors to debate whether or not a publishing company should invest in making a new game. A weakness is that it is only a prediction, and does not take into account things like the great recession.

Final Plots and Summary

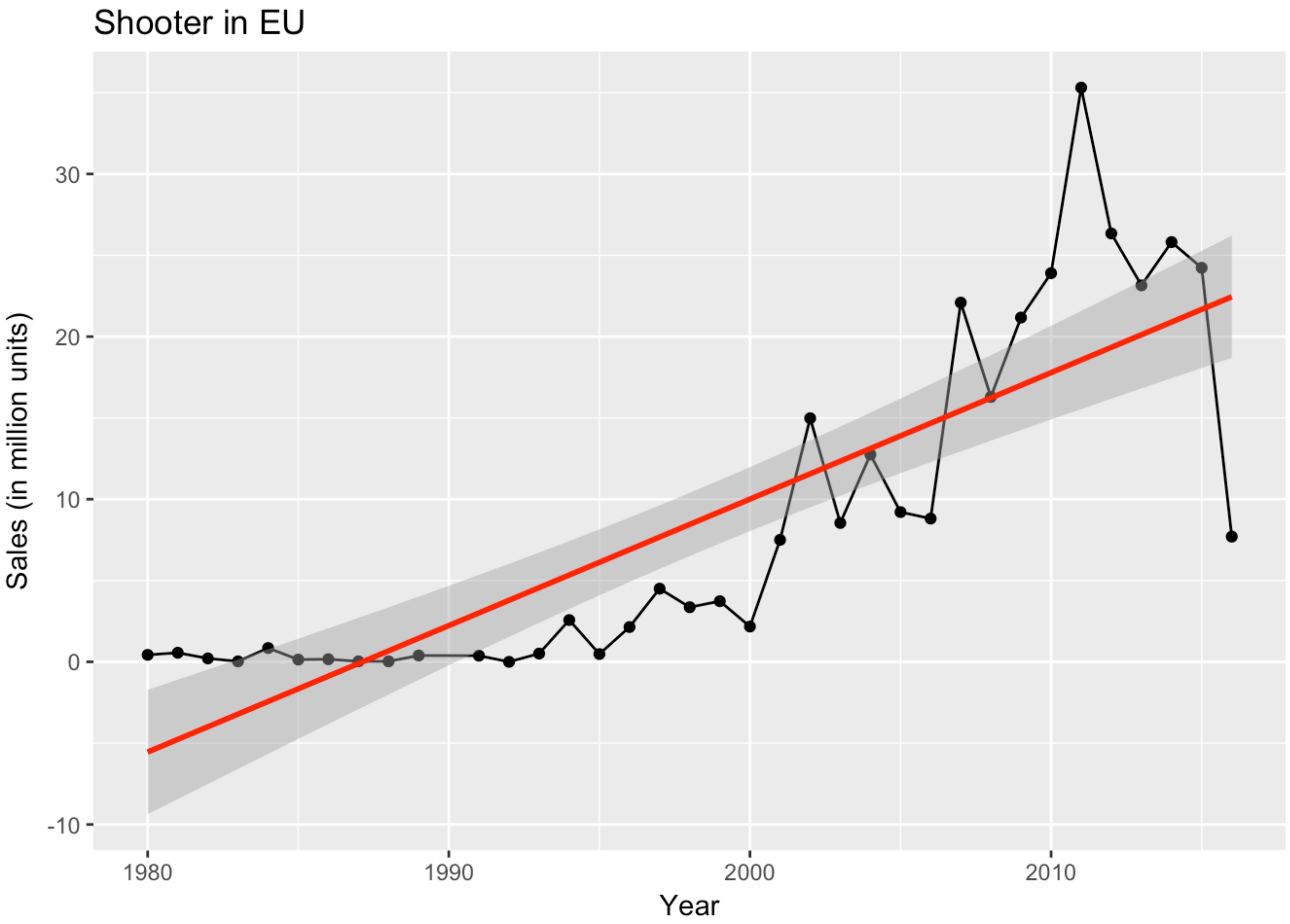
Plot One



Description One

I chose this graph because to me, it made the boldest statement. We can take this plot in the literal and see how sales increased then decreased, but when explore explanations in why, we paint a picture of real world events and how it affects the video game sales.

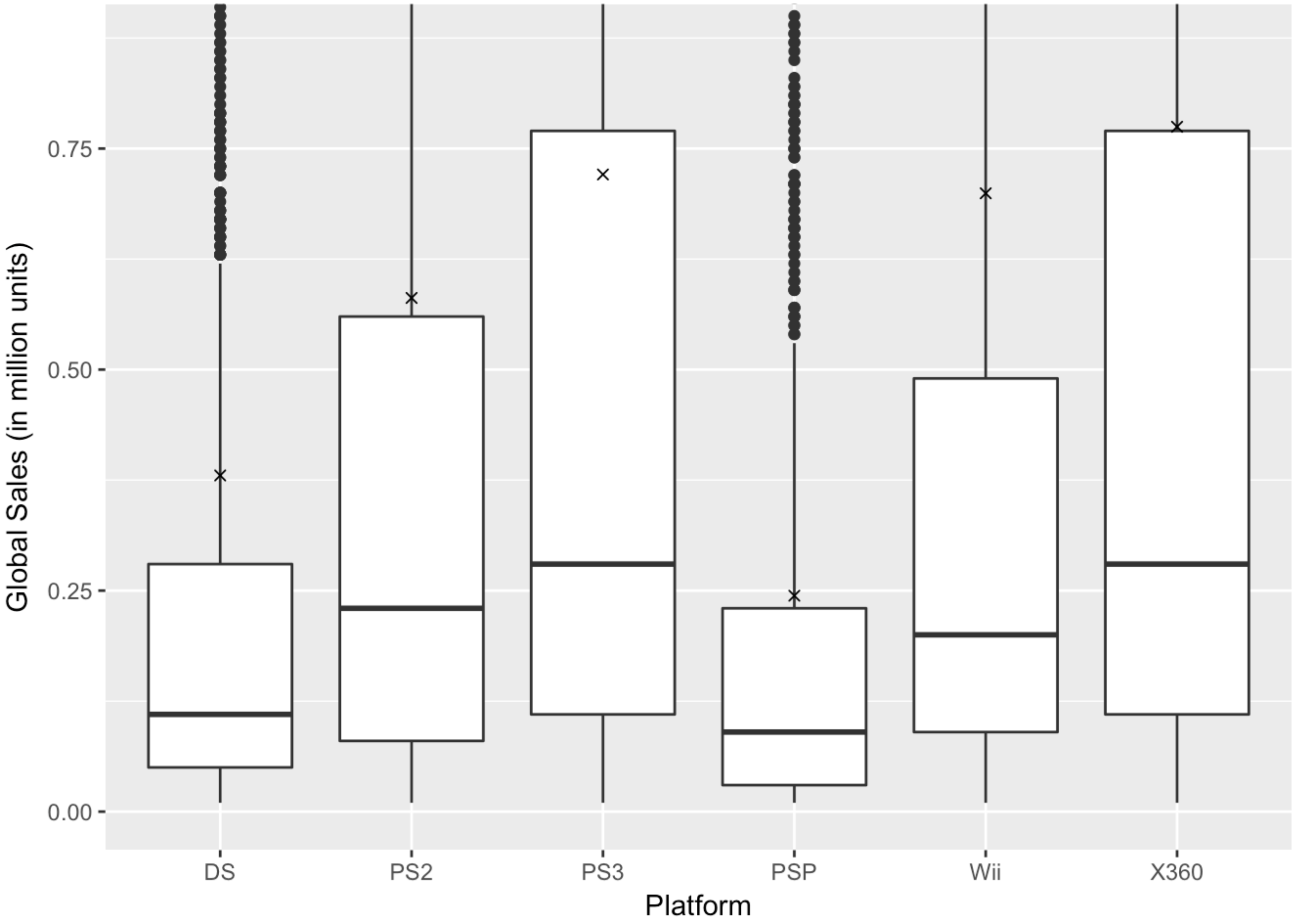
Plot Two



Description Two

I created this plot so we can see how shooters performed and predict how it will perform in the future. Another thing this plot shows is that in the EU, shooters did not decline until 2011, after the Great Recession. I chose this plot because while exploring the data for one thing, I found something unexpected.

Plot Three



Description Three

The “Console Wars” was about which consoles would be best to buy. This plot can tell us which consoles “won.” Of the three rival console, Wii, PS3, and Xbox360, Wii did not perform as well even though it had more games published. PS3 and Xbox360 did fairly similar.

Reflection

The video games data set contains sales data for video games created since 1980. By looking at my graphs i got a general idea how the history of game sales, and how worldly events (ei: new generations of consoles and the great recession). has an affect on the data. I would have like to further explore the data more by looking making more comparisions more sales regarding Genre and Region.

What limits the data is how the data was gathered. Looking at the VG Chartz (the site that provided the data), we can see the type of sales and a majority of the data were based on retail sales. After some thought, I came to realize that VG Chartz's data was incomplete in three main ways.

1. There was insufficient data regarding digital sales By providing games digitally, publishing companies can save more money and customers can easily obtain the games more conveniently VS retail. Although this data shows a decline in video games sales, since most of the data are about retail sales, I question whether the decline is actually that steep.
2. Indie Games/Mobile Games Most of the data are produced by big name publishers. However, producing a game through a publisher is not required. Many people can develop a game and sell it through other means such as steam.
3. In game purchases This probably has the biggest impact on our data. Games are now able to have extra download content and/or in-game purchases. So I can buy BioShock Infinite once and continue to purchase more content for the game later on. The data only reflects the purchase of the initial game and does not account for addition content for the game.

I struggled to reshape the data in order to create the desired plot. I had a lot of trouble figuring out how to reshape my data. Wide or Long, grouping vs aggregate vs gather(). Figuring out this process really made me think about how to organize my data, and how I could alter it make my plots.

improving the data, How can the data be improved? First and formost, not have data from 2020 or include presale fix the 271 data with no year Digital Sales VS Retail (not all included?) population of region (better to ratio it out?)

Reference

Data Kaggle url - https://www.kaggle.com/gregorut/videogamesales (https://www.kaggle.com/gregorut/videogamesales) VG Chartz - http://www.vgchartz.com/ (http://www.vgchartz.com/)

Udacity Data Analyst Courses - https://Udacity.com/ (https://Udacity.com/) Example Project - https://s3.amazonaws.com/content.udacity-data.com/courses/ud651/diamondsExample_2016-05.html (https://s3.amazonaws.com/content.udacity-data.com/courses/ud651/diamondsExample_2016-05.html) Fourms - https://discussions.udacity.com/c/nd002-data-analysis-with-r (https://discussions.udacity.com/c/nd002-data-analysis-with-r)

Questions and Queries - https://stackoverflow.com/ (https://stackoverflow.com/)