Video Game Sales Analysis

Hariprabhaa Murugesan

Introduction

The video games industry is worth billions of dollars, with companies spending vast amounts of money on the development and marketing of these games to an equally large market. It grows ever larger each year bringing new blockbuster hits. But all have achieved tens or even hundreds of millions in sales, and are unlikely to be unseated any time in the near future.Nintendo has made a lot of money by turning the Switch into a platform for independent developers, while Microsoft is pivoting to games as a service and Sony is continuing to move forward with an impressive line-up of exclusive titles.

Problem Identification

Looking for the best-selling games of all time, it's a mix of classics and current-generation staples. Some have stood the test of time, others very much seem to be products of their era. Through analysis of factors like most popular Genre, Platform with more number of games, country that contributes more to the sales , valuable insights can be drawn which helps in the development of video games industry and sales can be increased.

Load the required Packages

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------
------------------------------------------ tidyverse 1.2.1 --

## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------------------------
------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
```

Data Introduction

Dataset analyzed was generated by a scrape of vgchartz.com and contains a list of video games with sales greater than 100,000 copies from the Year 1980 to 2016.

Fields include:

Rank - Ranking of overall sales

Name - The games name

Platform - Platform of the games release (i.e. PC,PS4, etc.)

Year - Year of the game's release

Genre - Genre of the game

Publisher - Publisher of the game

NA_Sales - Sales in North America (in millions)

EU_Sales - Sales in Europe (in millions)

JP_Sales - Sales in Japan (in millions)

Other_Sales - Sales in the rest of the world (in millions)

Global_Sales - Total worldwide sales.

Load the data

```
vg<-read.csv(file = "C:/Users/omkum/Downloads/vgsales.csv ",header =
TRUE,stringsAsFactors = FALSE)
```

Data Cleaning

Before analysing the data, it needs to be cleaned. Nearly 300 NULL value records are removed from the dataset and Year is converted from character to numeric.

```
vg$Year<-suppressWarnings(as.numeric(vg$Year)) # convert character to numeric
vg<- na.omit(vg)
```

Data Overview

Top 5 records in the dataset

```
head(vg)
```

```
##   Rank                     Name Platform Year        Genre Publisher
## 1    1               Wii Sports      Wii 2006       Sports  Nintendo
```

```
## 2    2        Super Mario Bros.      NES 1985     Platform  Nintendo
## 3    3           Mario Kart Wii      Wii 2008       Racing  Nintendo
## 4    4        Wii Sports Resort      Wii 2009       Sports  Nintendo
## 5    5 Pokemon Red/Pokemon Blue       GB 1996 Role-Playing  Nintendo
## 6    6                   Tetris       GB 1989       Puzzle  Nintendo
##   NA_Sales EU_Sales JP_Sales Other_Sales Global_Sales
## 1    41.49    29.02     3.77        8.46        82.74
## 2    29.08     3.58     6.81        0.77        40.24
## 3    15.85    12.88     3.79        3.31        35.82
## 4    15.75    11.01     3.28        2.96        33.00
## 5    11.27     8.89    10.22        1.00        31.37
## 6    23.20     2.26     4.22        0.58        30.26
```

Last 5 records in the dataset

```
tail(vg)
```

```
##        Rank                                            Name Platform Year
## 16593 16595                                       Plushees       DS 2008
## 16594 16596              Woody Woodpecker in Crazy Castle 5      GBA 2002
## 16595 16597                    Men in Black II: Alien Escape       GC 2003
## 16596 16598 SCORE International Baja 1000: The Official Game      PS2 2008
## 16597 16599                                     Know How 2       DS 2010
## 16598 16600                                Spirits & Spells      GBA 2003
##            Genre  Publisher NA_Sales EU_Sales JP_Sales Other_Sales
## 16593 Simulation  Destineer     0.01     0.00        0           0
## 16594   Platform      Kemco     0.01     0.00        0           0
## 16595    Shooter Infogrames     0.01     0.00        0           0
## 16596     Racing Activision     0.00     0.00        0           0
## 16597     Puzzle   7G//AMES     0.00     0.01        0           0
## 16598   Platform    Wanadoo     0.01     0.00        0           0
##       Global_Sales
## 16593         0.01
## 16594         0.01
## 16595         0.01
## 16596         0.01
## 16597         0.01
## 16598         0.01
```

Structure of the dataset

```
str(vg)
```

```
## 'data.frame':    16327 obs. of  11 variables:
##  $ Rank        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Name        : chr  "Wii Sports" "Super Mario Bros." "Mario Kart Wii"
"Wii Sports Resort" ...
##  $ Platform    : chr  "Wii" "NES" "Wii" "Wii" ...
##  $ Year        : num  2006 1985 2008 2009 1996 ...
##  $ Genre       : chr  "Sports" "Platform" "Racing" "Sports" ...
##  $ Publisher   : chr  "Nintendo" "Nintendo" "Nintendo" "Nintendo" ...
```

```
##  $ 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 ...
##  - attr(*, "na.action")= 'omit' Named int  180 378 432 471 608 625 650 653
712 783 ...
##   ..- attr(*, "names")= chr  "180" "378" "432" "471" ...
```

Dimensions of the dataset

```
dim(vg)
```

```
## [1] 16327    11
```

Dataset Summary

```
summary(vg)
```

```
##       Rank           Name              Platform              Year
##  Min.   :    1   Length:16327       Length:16327       Min.   :1980
##  1st Qu.: 4136   Class :character   Class :character   1st Qu.:2003
##  Median : 8295   Mode  :character   Mode  :character   Median :2007
##  Mean   : 8293                                         Mean   :2006
##  3rd Qu.:12442                                         3rd Qu.:2010
##  Max.   :16600                                         Max.   :2020
##     Genre            Publisher            NA_Sales          EU_Sales
##  Length:16327       Length:16327       Min.   : 0.0000   Min.   : 0.0000
##  Class :character   Class :character   1st Qu.: 0.0000   1st Qu.: 0.0000
##  Mode  :character   Mode  :character   Median : 0.0800   Median : 0.0200
##                                        Mean   : 0.2654   Mean   : 0.1476
##                                        3rd Qu.: 0.2400   3rd Qu.: 0.1100
##                                        Max.   :41.4900   Max.   :29.0200
##     JP_Sales         Other_Sales        Global_Sales
##  Min.   : 0.00000   Min.   : 0.00000   Min.   : 0.0100
##  1st Qu.: 0.00000   1st Qu.: 0.00000   1st Qu.: 0.0600
##  Median : 0.00000   Median : 0.01000   Median : 0.1700
##  Mean   : 0.07866   Mean   : 0.04832   Mean   : 0.5402
##  3rd Qu.: 0.04000   3rd Qu.: 0.04000   3rd Qu.: 0.4800
##  Max.   :10.22000   Max.   :10.57000   Max.   :82.7400
```

Column Names in the datset

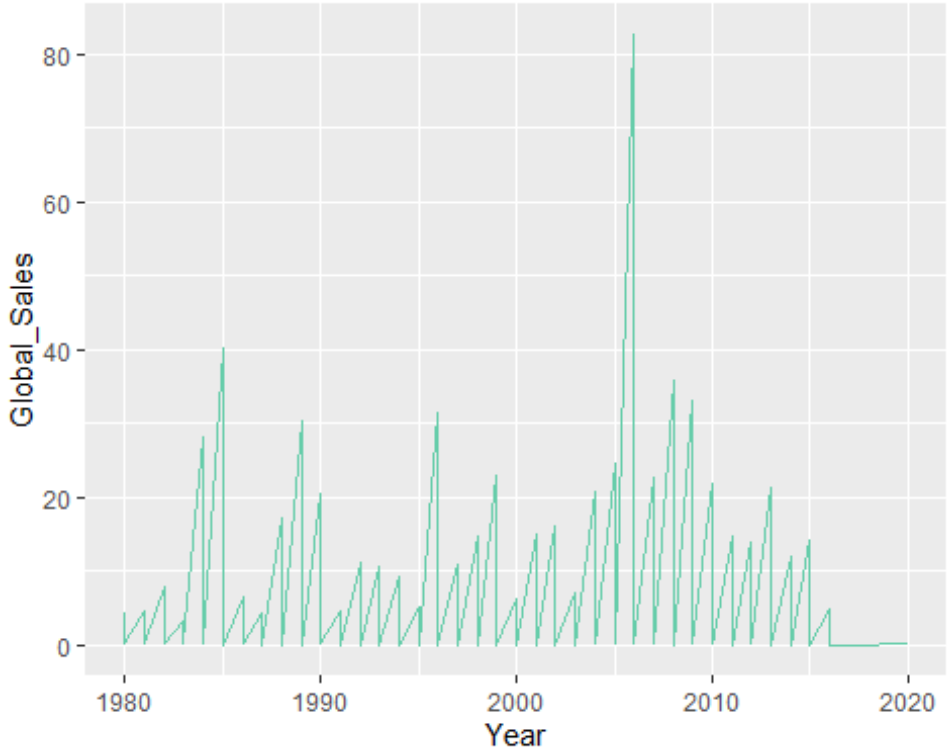
```
colnames(vg)
```

```
##  [1] "Rank"         "Name"         "Platform"     "Year"
##  [5] "Genre"        "Publisher"    "NA_Sales"     "EU_Sales"
##  [9] "JP_Sales"     "Other_Sales"  "Global_Sales"
```

Global Sales over the years

As we can see the video game sales was at its peak around 2005 and it went on decreasing over the course of years.

```
ggplot(vg,aes(x=Year,y =Global_Sales)) +geom_line(color="aquamarine3")
```



Analysis by Genre

The below bar graph depicts the number of games in each genre. Action genre contains more number of games followed by Sports, Role-playing and puzzle has the least count.

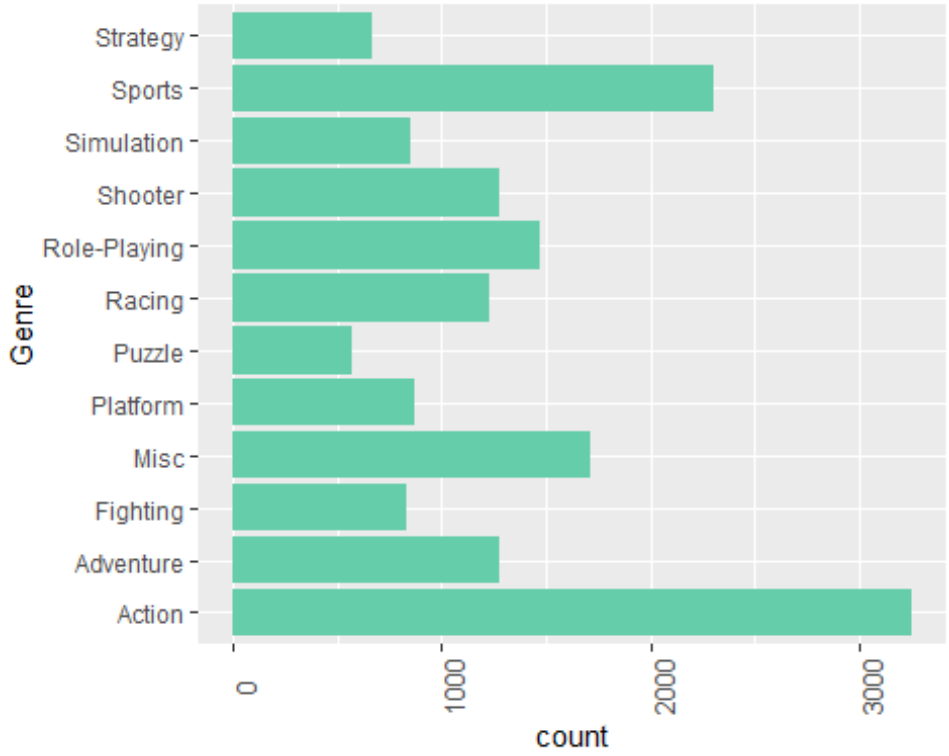
Count values are

```
vg %>%
  group_by(Genre) %>%
  count(Genre) %>%
  arrange(desc(n))

## # A tibble: 12 x 2
## # Groups:   Genre [12]
##    Genre            n
##    <chr>        <int>
##  1 Action        3253
##  2 Sports        2304
##  3 Misc          1710
```

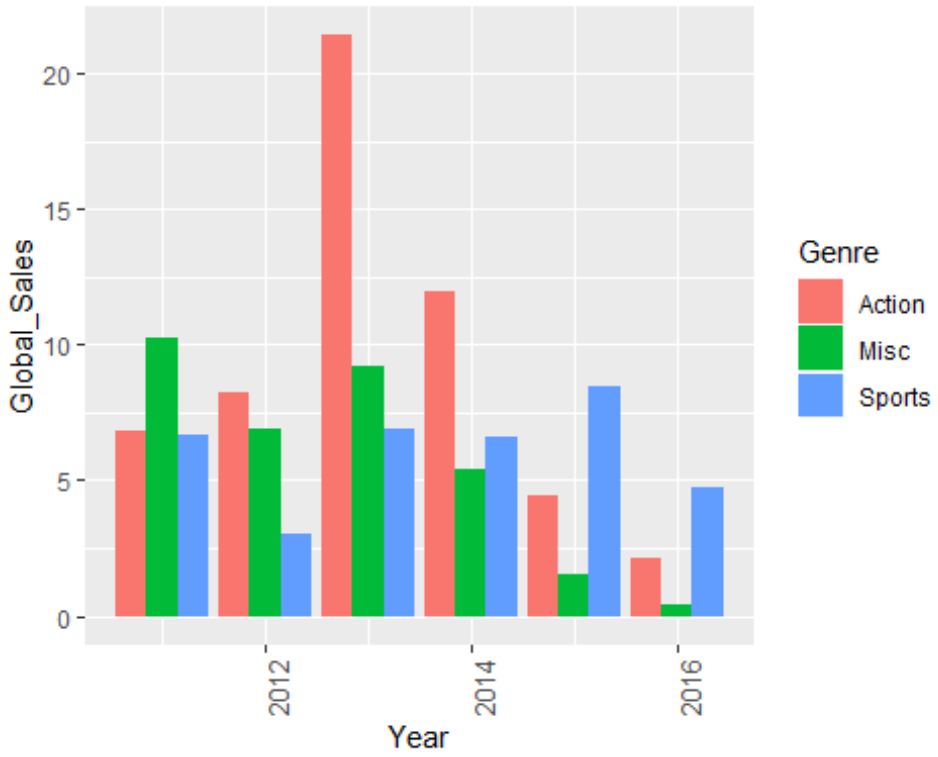
```
##  4 Role-Playing  1471
##  5 Shooter       1282
##  6 Adventure     1276
##  7 Racing        1226
##  8 Platform       876
##  9 Simulation     851
## 10 Fighting       836
## 11 Strategy       671
## 12 Puzzle         571
```

```
ggplot(data=vg, aes(x=Genre)) +
    geom_bar(stat="count", fill= "aquamarine3")+theme(axis.text.x =
element_text(size=10, angle=90))+coord_flip()
```



Since most of the games belongs to Action, Sports and Misc Genre, stacked bar has been plotted for the Global Sales accross these genres each year. Action games were sold out in more numbers before 2015 and Sports genre is dominating later.

```
vg %>%
  filter((Genre=="Action" | Genre=="Sports" | Genre=="Misc") & Year> 2010 &
Year<2017) %>%
  ggplot(aes(x=Year,y=Global_Sales,fill=Genre)) +geom_bar(stat = "identity" ,
position=position_dodge())+theme(axis.text.x =   element_text(size=10,
angle=90))
```



Analysis by platform

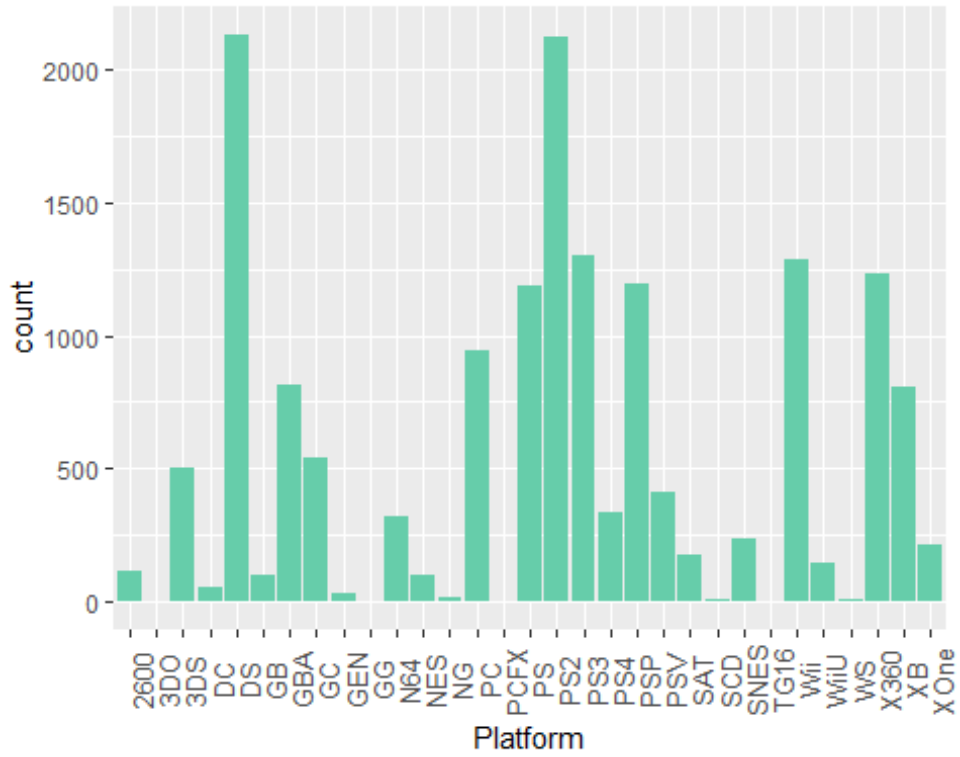
The below bar graph depicts the number of games in each Platform. Ds tops the table with 2,163 games followed by PS2, PS3.

Count values are

```
vg %>%
  group_by(Platform) %>%
  count(Platform) %>% arrange(desc(n))
```

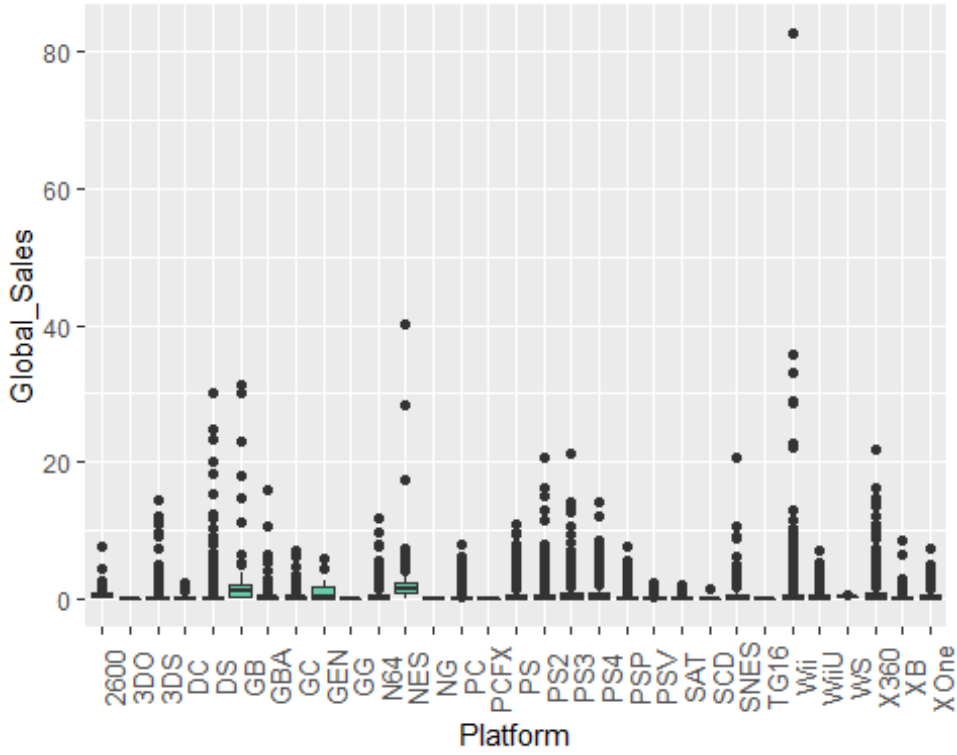
```
## # A tibble: 31 x 2
## # Groups:   Platform [31]
##    Platform     n
##    <chr>    <int>
##  1 DS        2133
##  2 PS2       2127
##  3 PS3       1304
##  4 Wii       1290
##  5 X360      1235
##  6 PSP       1197
##  7 PS        1189
##  8 PC         943
##  9 GBA        811
## 10 XB         803
## # ... with 21 more rows
```

```
ggplot(data=vg, aes(x=Platform)) +
    geom_bar(stat="count",fill="aquamarine3")+theme(axis.text.x =
element_text(size=10, angle=90))
```



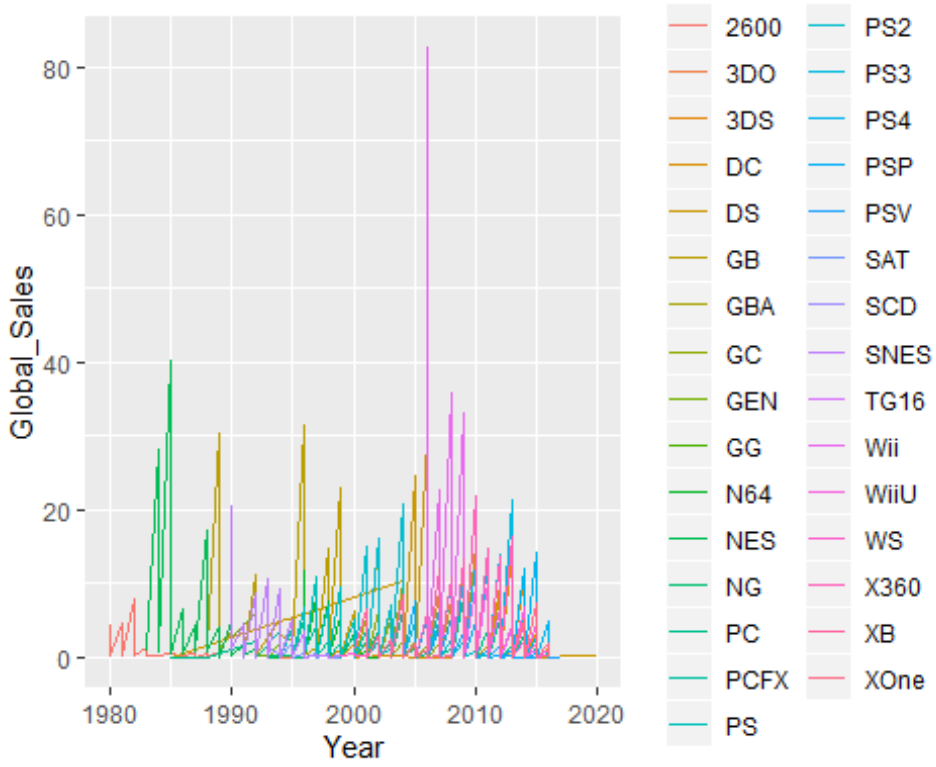
Boxplot indicates the distribution of sales based on platform.

```
ggplot(vg,aes(x=Platform,y=Global_Sales))+geom_boxplot(fill="aquamarine3") +
theme(axis.text.x = element_text(size=10, angle=90))
```



Sales by Platform each year has been shown in the below graph.

```
ggplot(vg,aes(x=Year,y=Global_Sales,group=Platform, color=Platform))
+geom_line()
```



Region wise Sales Analysis

Analysing the maximum sales in each region, Wii Sports takes the lead position in both North America & Europe whereas Pokemon tops the list in Japan and Grand Theft in other countries.

```
vg[which.max(vg$NA_Sales),2]
```

```
## [1] "Wii Sports"
```

```
vg[which.max(vg$EU_Sales),2]
```

```
## [1] "Wii Sports"
```

```
vg[which.max(vg$JP_Sales),2]
```

```
## [1] "Pokemon Red/Pokemon Blue"
```

```
vg[which.max(vg$Other_Sales),2]
```

```
## [1] "Grand Theft Auto: San Andreas"
```

Below table contains Region wise Sales each year. Japan have no video games sales till 1983 whereas North America has good sales over the years.

```
vg %>%
  group_by(Year) %>%
```

```
summarise_at(vars("JP_Sales","NA_Sales","EU_Sales","Other_Sales","Global_Sales"),sum)
```

```
## # A tibble: 39 x 6
##     Year JP_Sales NA_Sales EU_Sales Other_Sales Global_Sales
##    <dbl>    <dbl>    <dbl>    <dbl>       <dbl>        <dbl>
##  1  1980      0      10.6      0.67        0.12         11.4
##  2  1981      0      33.4      1.96        0.32         35.8
##  3  1982      0      26.9      1.65        0.31         28.9
##  4  1983      8.1     7.76     0.8         0.14         16.8
##  5  1984     14.3    33.3      2.1         0.7          50.4
##  6  1985     14.6    33.7      4.74        0.92         53.9
##  7  1986     19.8    12.5      2.84        1.93         37.1
##  8  1987     11.6     8.46     1.41        0.2          21.7
##  9  1988     15.8    23.9      6.59        0.99         47.2
## 10  1989     18.4    45.2      8.44        1.5          73.4
## # ... with 29 more rows
```

Analysis by Publisher

Publisher with maximum sales:

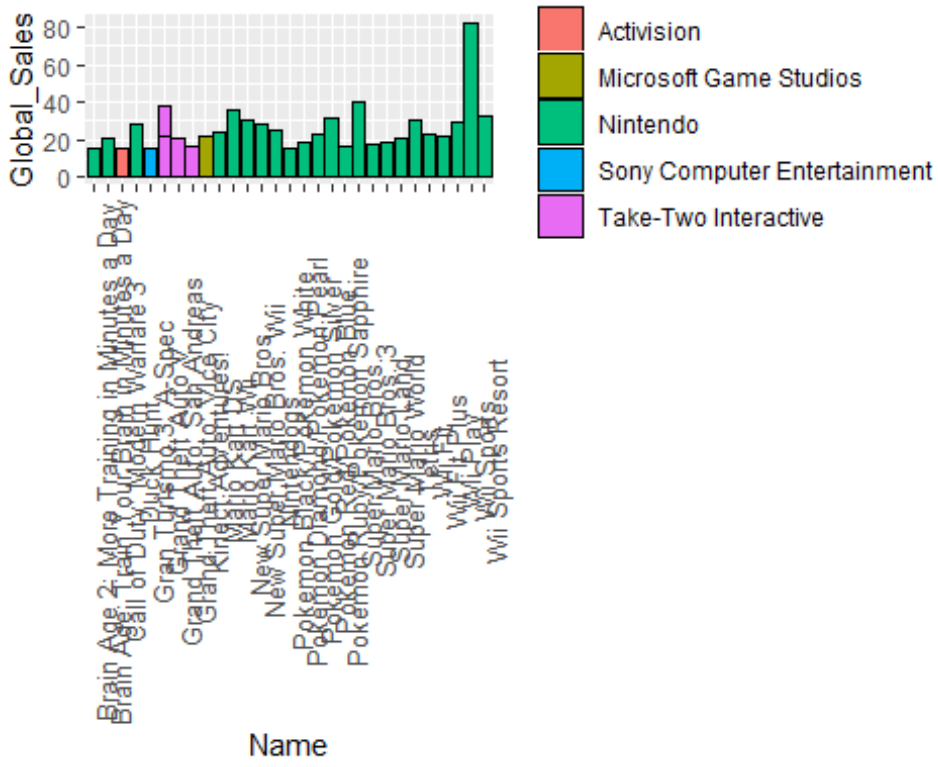
```
x<-vg %>%
  group_by(Publisher) %>%
  summarise(sum(Global_Sales))

vg[which.max(x$`sum(Global_Sales)`),]
```

```
##     Rank        Name Platform Year   Genre       Publisher NA_Sales
## 370  370 Left 4 Dead     X360 2008 Shooter Electronic Arts     2.66
##     EU_Sales JP_Sales Other_Sales Global_Sales
## 370      0.5     0.05         0.3         3.52
```

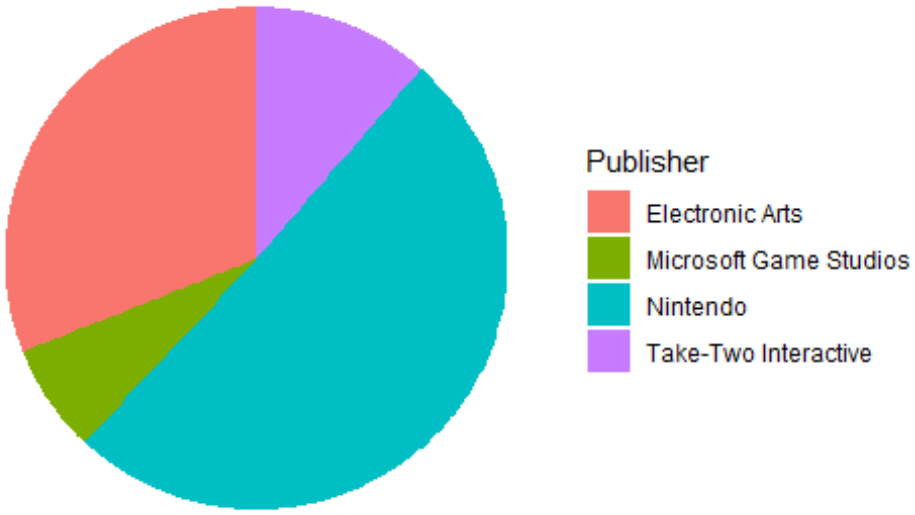
Top selling 20 games in the market and the corresponding publishers. As per the graph, most of the games belongs to Nintendo.

```
vg  %>%
  arrange(desc(Global_Sales)) %>%
  head(30)  %>%ggplot(aes(x=Name,y=Global_Sales,fill=Publisher))
+geom_bar(stat = "identity",colour="black") +theme(axis.text.x =
element_text(size=10, angle=90))
```



Pie chart demonstrates the Global Sales among the top 4 popular Publishers.

```
vg %>%
  filter(Publisher=="Nintendo" | Publisher=="Electronic Arts" |
Publisher=="Take-Two Interactive" | Publisher=="Microsoft Game Studios") %>%
  ggplot(aes(x="", y=Global_Sales, fill=Publisher)) +
  geom_bar(stat="identity", width=1) +
  coord_polar("y", start=0) +
  theme_void()
```



Prediction of Global sales

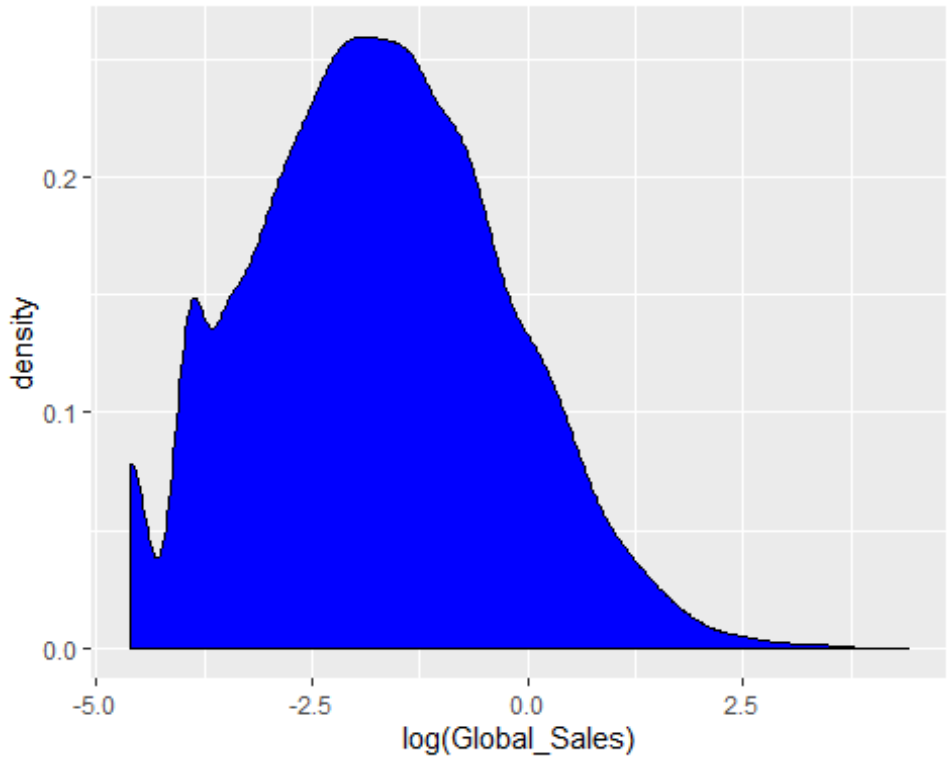
On trying to found out the region that contributes more towards the global sales,correlation matrix is constructed. As per the coefficient, North America influences the most.

```
cor(vg[sapply(vg, is.numeric)])
```

```
##                    Rank         Year    NA_Sales     EU_Sales   JP_Sales
## Rank          1.0000000  0.178813640 -0.40032832 -0.379023676 -0.2691378
## Year          0.1788136  1.000000000 -0.09140216  0.006013887 -0.1693162
## NA_Sales     -0.4003283 -0.091402162  1.00000000  0.768936262  0.4512854
## EU_Sales     -0.3790237  0.006013887  0.76893626  1.000000000  0.4364139
## JP_Sales     -0.2691378 -0.169316218  0.45128538  0.436413946  1.0000000
## Other_Sales  -0.3325212  0.041057667  0.63450832  0.726265653  0.2906527
## Global_Sales -0.4268798 -0.074734798  0.94126766  0.903270981  0.6127938
##              Other_Sales Global_Sales
## Rank          -0.33252121   -0.4268798
## Year           0.04105767   -0.0747348
## NA_Sales       0.63450832    0.9412677
## EU_Sales       0.72626565    0.9032710
## JP_Sales       0.29065268    0.6127938
## Other_Sales    1.00000000    0.7479742
## Global_Sales   0.74797420    1.0000000
```

NOrmal Distribution for log(global sales).

```
ggplot(vg, aes(log(Global_Sales))) + geom_density(fill="blue")
```



Split the data for training and testing.

```
set.seed(1)
row.number <- sample(1:nrow(vg), 0.8*nrow(vg))
train = vg[row.number,]
test = vg[-row.number,]
dim(train)

## [1] 13061    11

dim(test)

## [1] 3266   11
```

Below information contains the summary of model constructed.

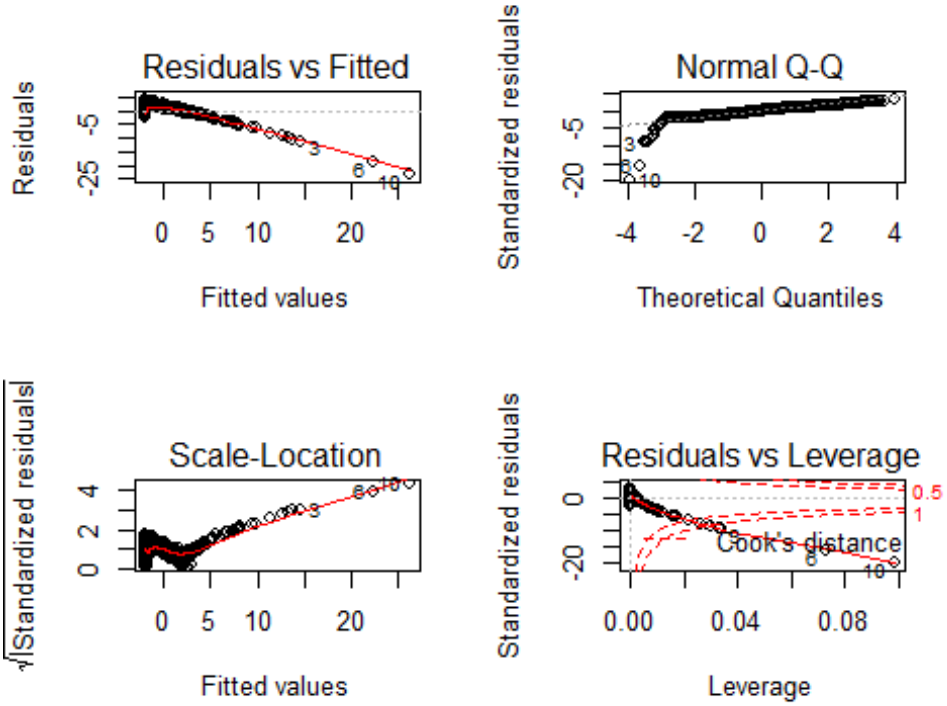
```
model1 = lm(log(Global_Sales)~NA_Sales, data=train)
summary(model1)

##
## Call:
## lm(formula = log(Global_Sales) ~ NA_Sales, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.9551  -0.8286   0.1228   0.9376   3.7801
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.01636    0.01149 -175.49   <2e-16 ***
## NA_Sales     1.05142    0.01459   72.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.239 on 13059 degrees of freedom
## Multiple R-squared:  0.2847, Adjusted R-squared:  0.2846
## F-statistic:  5196 on 1 and 13059 DF,  p-value: < 2.2e-16
```

Plot the model.

```
par(mfrow=c(2,2))
plot(model1)
```



On predicting the Global sales, actual and predicted values are as follows:

```
globsalespredict <- predict(model1, test)
actuals_preds <- data.frame(cbind(actuals=test$Global_Sales,
predicteds=globsalespredict))
head(actuals_preds)
```

```
##    actuals predicteds
## 1    82.74  41.606902
## 2    40.24  28.558824
## 11   24.76   7.519983
```

```
## 12   23.42   8.298031
## 17   21.40   5.354065
## 25   16.15   6.826048
```

Conclusion

Based on the analysis and plots of sales data, the following conclusions can be drawn:

- Sports genre is gaining more popularity in recent times and hence development of this type of games will increase the sales. But Japan shows more interest in role playing games. Interest towards Action genre games is decreasing day by day.
- Games published by Nintendo gets more attention. Still Sony, Electronic Arts Sustains some top position in Global sales.
- North America contributes more towards the Global Sales based on the Correlation matrix. Through increasing sales in this region, game industry might gain some profit. At the same time, marketing can be increased in rest of the regions.