Práctica 2: Limpieza y validación de los datos

Byron Vinicio Lima Rojas

28 de diciembre, 2017

Contents

Práctica 2

El objetivo de esta actividad será el tratamiento de un dataset, que puede ser el creado en la práctica 1 o cualquier dataset libre disponible en Kaggle (https://www.kaggle.com).

Para esta práctica utilizaremos un dataset de Kaggle que contiene información de los 16.598 videojuegos con una fecha de inicio del año 1980, así como sus datos de ventas en un total de 11 variables.

```
videogames<-read.csv("vgsales.csv", sep=",",na.strings = "NA", dec = ".")
```

Las variables son:

- **Rank** - Ranking de las ventas totales.
- **Name** - El nombre de los juegos.
- **Platform** - Plataforma de lanzamiento de juegos
- **Year** - Año del lanzamiento
- **Genre** - Género
- **Publisher** - Editor
- **NA_Sales** - Ventas en América del Norte (millones)
- **EU_Sales** - Ventas en Europa (millones)
- **JP_Sales** - Ventas en Japón (millones)
- **Other_Sales** - Ventas en el resto del mundo (millones)
- **Global_Sales** - Total de ventas en todo el mundo.

1. Descripción del Dataset

La industria de los videojuegos siempre ha sido rentable, siendo una de las principales industrias del arte y entretenimiento ya que constantemente la mayoría de productoras de películas y generadoras de contenido multimedia están buscando posesionarse en lo más alto en el mercado actual, cuando un producto inicial es atractivo al público buscan crear nuevo contenido a partir del mismo (series, revistas, videojuegos).

Sin embargo, dar con la idea de qué tipo de videojuego crear y a que empresa encomendar dicha tarea no es fácil, es por cuanto el Dataset que va a ser analizado contiene datos de videojuegos desde el año 1980, en donde se revisaran los editores con más éxito en la venta de videojuegos y nos permitirá conocer qué tipo de videojuego es más rentable, de esta forma podremos brindar recomendaciones a nuevas compañías en cuando a que contenido crear, tipo de historia, la clasificación que debe tener el videojuego y a que plataforma enfocarse para ser un producto exitoso.

```
# CORRECCIÓN DE SEPARADORES DECIMALES
# Cuando se grabe como .csv2 quedará como , decimal.
videogames$NA_Sales <-  as.numeric(sub(",","\\.",videogames$NA_Sales))
videogames$EU_Sales <-  as.numeric(sub(",","\\.",videogames$EU_Sales))
videogames$JP_Sales <-  as.numeric(sub(",","\\.",videogames$JP_Sales))
videogames$Other_Sales <-  as.numeric(sub(",","\\.",videogames$Other_Sales))
videogames$Global_Sales <-  as.numeric(sub(",","\\.",videogames$Global_Sales))
# Descripción de variables contenidas en el Dataset.
str(videogames)
```

```
## 'data.frame':    16598 obs. of  11 variables:
##  $ Rank        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Name        : Factor w/ 11493 levels "'98 Koshien",..: 10991 9343 5532 10993 7370 9707 6648 10989
##  $ 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 ...
```

2. Limpieza de los datos.

2.1. Selección de los datos de interés a analizar. ¿Cuáles son los campos más relevantes para responder al problema?

De acuerdo al objetivo inicial del uso de estos datos, los campos más relevantes para realizar una propuesta son los siguientes:

- **Rank** - Ranking de las ventas totales.
- **Name** - El nombre de los juegos.
- **Platform** - Plataforma de lanzamiento de juegos
- **Year** - Año del lanzamiento
- **Genre** - Género
- **Publisher** - Editor
- **Global_Sales** - Total de ventas en todo el mundo.

De estos datos es importante analizarlos por separado, de acuerdo al ranking de ventas, año de publicación género y ventas globales. Estos datos serán analizados por separados para realizar propuestas significativas.

2.2. ¿Los datos contienen ceros o elementos vacíos? ¿Y valores extremos? ¿Cómo gestionarías cada uno de estos casos?

De las variables seleccionadas, procederemos a revisar los datos que contiene cada una:

```
table(videogames$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
```

```
table(videogames$Year)
```

```
##
## 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994
##    9   46   36   17   14   14   21   16   15   17   16   41   43   60  121
## 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009
##  219  263  289  379  338  349  482  829  775  763  941 1008 1202 1428 1431
## 2010 2011 2012 2013 2014 2015 2016 2017 2020  N/A
## 1259 1139  657  546  582  614  344    3    1  271
```

```
table(videogames$Genre)
```

```
##
##       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
```

De los datos del editor se revisaron manualmente por la cantidad de datos y no existen datos que contienen ceros o elementos vacíos.

De acuerdo a lo revisado, en años ya hay un valor atípico con el dato en año 2020 y hay videojuegos que no tienen fecha de publicación, para este caso podemos imputar los valores a partir de los k-vecinos más próximos o eliminarlos. Al ser únicamente 278 registros con este inconveniente no los tomaremos en cuenta.

```
#Revisión de datos en Ventas
summary(videogames$Global_Sales)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0100  0.0600  0.1700  0.5374  0.4700 82.7400
```

Claramente entre el cuarto y quinto quartil hay mucha diferencia, puesto que existen datos muy antiguos para ser analizados, los cuales al momento con los distintos cambios en la tecnología no sería prudente tomarlos en cuenta, adicional podría ser que en el análisis recomendemos un juego que fue diseñado para una plataforma no disponible en el mercado. En base a estas observaciones utilizaremos únicamente los videojuegos de los últimos 10 años (2007 - 2017) y de plataformas existentes.

```r
#Creación de un nuevo Dataset a partir del año.
vgamesP <- subset(videogames, videogames$Year == "2007"
                        | videogames$Year == "2008" | videogames$Year == "2009"
                        | videogames$Year == "2010" | videogames$Year == "2011"
                        | videogames$Year == "2012" | videogames$Year == "2013"
                        | videogames$Year == "2014" | videogames$Year == "2015"
                        | videogames$Year == "2016" | videogames$Year == "2017",
                       select = c(Rank, Name, Platform, Year, Genre, Publisher,
                                  Global_Sales))


#Creación de Dataset Final a partir
videogamesnew <- subset(vgamesP, vgamesP$Platform == "3DS"
                        | vgamesP$Platform == "DS" | vgamesP$Platform == "PC"
                        | vgamesP$Platform == "PS2" | vgamesP$Platform == "PS3"
                        | vgamesP$Platform == "PS4" | vgamesP$Platform == "PSP"
                        | vgamesP$Platform == "PSV" | vgamesP$Platform == "Wii"
                        | vgamesP$Platform == "X360" | vgamesP$Platform == "XOne")

#Detalle de datos
head(videogamesnew)
```

```
##    Rank                      Name Platform Year    Genre
## 3     3            Mario Kart Wii      Wii 2008   Racing
## 4     4         Wii Sports Resort      Wii 2009   Sports
## 9     9 New Super Mario Bros. Wii      Wii 2009 Platform
## 14   14                  Wii Fit      Wii 2007   Sports
## 15   15             Wii Fit Plus      Wii 2009   Sports
## 16   16        Kinect Adventures!     X360 2010     Misc
##                  Publisher Global_Sales
## 3                 Nintendo        35.82
## 4                 Nintendo        33.00
## 9                 Nintendo        28.62
## 14                Nintendo        22.72
## 15                Nintendo        22.00
## 16 Microsoft Game Studios        21.82
```

```r
#Número de Registros
nrow(videogamesnew)
```

```
## [1] 9046
```

```r
#Número de columnas
ncol(videogamesnew)
```

```
## [1] 7
```

3. Análisis de los datos.

3.1. Selección de los grupos de datos que se quieren analizar/comparar.

Después de realizar la limpieza de datos es importante detallar los procedimientos realizados para lograr un conjunto de datos limpio y adaptable a las conclusiones que queremos llegar:

- Exclusión de valores atípicos en los años del conjunto.
- Exclusión de registros con valores perdidos o campos vacíos puesto que los mismos no representaban un gran número de datos que impidan realizar el análisis de datos.
- Exclusión de registros de acuerdo a su monto de ventas, se realizó mediante la revisión de las distancias entre los cuartiles presentes en los valores de venta globales por cada uno de los registros, de esta forma todo videojuego con ventas globales menores a 0.75 millones / dólares no será tomado en cuenta.

Como uno de los propósitos es realizar una recomendación hacia la nueva industria sobre qué características debe tener un juego para ser exitoso, vamos a realizar una comparativa entre las siguientes variables existentes en el conjunto de datos:

- **Genre / Global_Sales:** Análisis del número de ventas en base al género
- **Platform / Global_Sales:** Análisis del número de ventas en base a la plataforma del videojuego.
- **Publisher / Global_Sales:** Análisis del número de ventas de cada uno de los editores.
- **Rank / Genre:** Análisis del género en base al ranking del Top 50 de videojuegos.
- **Rank / Platform:** Análisis de la plataforma en base al ranking del Top 50 de videojuegos.

A partir del análisis de estas variables, se podrá llegar a las conclusiones necesarias recomendar puntos importantes que se deben tomar en cuenta antes de crear un videojuego.

3.2. Comprobación de la normalidad y homogeneidad de la varianza. Si es necesario (y posible), aplicar transformaciones que normalicen los datos.

Al tener un solo dato numérico en la selección de datos a analizar, vamos a revisar determinadas características:

```
summary(videogamesnew$Global_Sales)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0100  0.0500  0.1500  0.4855  0.4300 35.8200
```

En base a los datos, del 3er Quartil (0,4300 millones) y el valor máximo (35.82 millones). Procedemos a reducir el número de registros en base a videojuegos que han realizado ventas mayores a 0.75 millones de copias.

```
videogamesnewS <- subset(videogamesnew, videogamesnew$Global_Sales > 0.75)
nrow(videogamesnewS)
```

```
## [1] 1334
```

Después de este filtro nos hemos quedado con 1334 registros en el Dataset **videogamesnewS**. A partir de los mismos, empezaremos a realizar las pruebas necesarias para llegar a una recomendación idónea. En base a este filtro, se puede ver que existen valores atípicos dentro de las ventas globales, pero eso únicamente determina el éxito en ventas del videojuego lo cual analizaremos en el siguiente enunciado mediante grupo de datos.

A continuación, se procede con la graficación de las tres variables más importantes en relación a su monto de ventas a nivel mundial:

Validación de variables normales con Shapiro

```
#Ventas globales
media <- mean(videogamesnewS$Global_Sales)
desviacionestandar <- sd(videogamesnewS$Global_Sales)
shapiro.test(videogamesnewS$Global_Sales)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  videogamesnewS$Global_Sales
```

```
## W = 0.49702, p-value < 2.2e-16
```

Interpretación: siendo la hipótesis nula que las ventas están distribuidas normalmente, en este caso p-valor es menor a alfa, entonces esta hipótesis nula es rechazada, concluyendo de esta forma que los datos de ventas globales no provienen de una distribución normal.

3.3. Aplicación de pruebas estadísticas (tantas como sea posible) para comparar los grupos de datos.

Antes de empezar las pruebas estadisticas, es importante conocer por separado que datos tenemos en cada grupo de datos

```
#Plataforma de videojuegos
table(videogamesnewS$Platform)
```

```
##
## 2600  3DO  3DS   DC   DS   GB  GBA   GC  GEN   GG  N64  NES   NG   PC PCFX
##    0    0   63    0  146    0    0    0    0    0    0    0    0   47    0
##   PS  PS2  PS3  PS4  PSP  PSV  SAT  SCD SNES TG16  Wii WiiU   WS X360   XB
##    0   48  327   88   45   13    0    0    0    0  210    0    0  296    0
## XOne
##   51
```

```
#Género
table(videogamesnewS$Genre)
```

```
##
##       Action    Adventure     Fighting         Misc     Platform
##          334           29           64          134           59
##       Puzzle       Racing Role-Playing      Shooter   Simulation
##           24           76          136          193           61
##       Sports     Strategy
##          208           16
```

```
#Año de venta
table(videogamesnewS$Year)
```

```
##
## 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994
##    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009
##    0    0    0    0    0    0    0    0    0    0    0    0  184  230  180
## 2010 2011 2012 2013 2014 2015 2016 2017 2020  N/A
##  165  149  121  109   98   76   22    0    0    0
```

Después de la revisión de datos individuales se puede evidenciar que hay plataformas de videojuego que ya no son utilizadas, adicional solo estamos utilizando datos a partir del 2008, de acuerdo a los géneros no existe inconsistencia en valores de variables. Para lo cual procedemos a realizar una tabla de los datos relevantes de nuestro subconjunto de datos:

```
library(knitr)
options(knitr.kable.NA = '')

#Revisión de datos del data set VideoGamesNewS.
kable(summary(videogamesnewS)[,c(3,4,5,6)],
      digits=2, caption="Estadística descriptiva de variables")
```

Table 1: Estadística descriptiva de variables

Platform	Year	Genre	Publisher
PS3 :327	2008 :230	Action :334	Electronic Arts :245
X360 :296	2007 :184	Sports :208	Ubisoft :126
Wii :210	2009 :180	Shooter :193	Activision :124
DS :146	2010 :165	Role-Playing:136	Nintendo :112
PS4 : 88	2011 :149	Misc :134	Take-Two Interactive : 80
3DS : 63	2012 :121	Racing : 76	Sony Computer Entertainment: 79
(Other):204	(Other):305	(Other) :253	(Other) :568

```
#Revisión del género y nombres de videojuegos del total de datos.
vgamesP <- subset(videogames, videogames$Year != "2020"
                          | videogames$Year == "NA",
                          select = c(Rank, Name, Platform, Year, Genre, Publisher,
                                     Global_Sales))
kable(summary(vgamesP)[,c(2,5)],
      digits=2)
```

Name	Genre
Need for Speed: Most Wanted: 12	Action :3316
FIFA 14 : 9	Sports :2346
LEGO Marvel Super Heroes : 9	Misc :1739
Madden NFL 07 : 9	Role-Playing:1488
Ratatouille : 9	Shooter :1310
Angry Birds Star Wars : 8	Adventure :1286
(Other) :16541	(Other) :5112

El objetivo de revisar los nombres de los videojuegos más vendidos y los géneros en sí, es para determinar qué tan alejados estamos de los datos que proponemos en nuestro subconjunto, hay que tomar en cuenta que las plataformas van evolucionando, pero los géneros de videojuegos y la forma de jugarlos permanecen.

Bueno, ya tenemos datos de las plataformas, géneros y editores con mayor número de publicaciones, es necesario conocer la cantidad de copias que consiguieron vender, para fundamentar nuestros resultados finales, para ello procederemos con la creación de suconjuntos.

Ventas globales por plataforma

```
#Plataforma PS3
PS301 <- subset(videogamesnewS, videogamesnewS$Platform == "PS3")
PS3 <- sum (PS301$Global_Sales)
PS3
```

```
## [1] 699.74
```

```
#Plataforma X360
X36001 <- subset(videogamesnewS, videogamesnewS$Platform == "X360")
X360 <- sum (X36001$Global_Sales)
X360
```

```
## [1] 711.19
```

```
#Plataforma Wii
Wii01 <- subset(videogamesnewS, videogamesnewS$Platform == "Wii")
```

```
Wii <- sum (Wii01$Global_Sales)
Wii
```

```
## [1] 553.91
```

```
#Plataforma DS
DS01 <- subset(videogamesnewS, videogamesnewS$Platform == "DS")
DS <- sum (DS01$Global_Sales)
DS
```

```
## [1] 294.39
```

```
#Plataforma PS4
PS401 <- subset(videogamesnewS, videogamesnewS$Platform == "PS4")
PS4 <- sum (PS401$Global_Sales)
PS4
```

```
## [1] 232.28
```

```
#Plataforma 3DS
EDS01 <- subset(videogamesnewS, videogamesnewS$Platform == "3DS")
EDS <- sum (EDS01$Global_Sales)
EDS
```

```
## [1] 175.37
```

```
#Plataforma PC
PC01 <- subset(videogamesnewS, videogamesnewS$Platform == "PC")
PC <- sum (PC01$Global_Sales)
PC
```

```
## [1] 83.43
```

```
#Plataforma PS2
PS201 <- subset(videogamesnewS, videogamesnewS$Platform == "PS2")
PS2 <- sum (PS201$Global_Sales)
PS2
```

```
## [1] 75.56
```

```
#Plataforma XOne
XOne01 <- subset(videogamesnewS, videogamesnewS$Platform == "XOne")
XOne <- sum (XOne01$Global_Sales)
XOne
```

```
## [1] 105.76
```

```
#Análisis de 5-números de 5 plataformas de videojuegos más vendidos
summary (c(PS3,X360,Wii,DS,PS4))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   232.3   294.4   553.9   498.3   699.7   711.2
```

Conclusión: De los 1334 registros de nuestro subconjunto estamos revisando únicamente las plataformas con mejores ventas, a esto hay que considerar que un videojuego está disponible para varias versiones, lo cual sustentaremos en base a la revisión por género de venta. De los valores de ventas, están cerca del valor máximo del total de datos. Para la revisión del género, analizaremos datos desde el 2008 realizando una comparativa de 5 números a partir de los 5 mejores géneros.

Ventas globales por Género

```r
#Género Action
Action01 <- subset(videogamesnewS, videogamesnewS$Genre == "Action")
Action <- sum (Action01$Global_Sales)
Action
```

```
## [1] 689.95
```

```r
#Género Sports
Sports01 <- subset(videogamesnewS, videogamesnewS$Genre == "Sports")
Sports <- sum (Sports01$Global_Sales)
Sports
```

```
## [1] 461.33
```

```r
#Género Shooter
Shooter01 <- subset(videogamesnewS, videogamesnewS$Genre == "Shooter")
Shooter <- sum (Shooter01$Global_Sales)
Shooter
```

```
## [1] 555.13
```

```r
#Género Role-Playing
RolePlaying01 <- subset(videogamesnewS, videogamesnewS$Genre == "Role-Playing")
RolePlaying <- sum (RolePlaying01$Global_Sales)
RolePlaying
```

```
## [1] 324.66
```

```r
#Género Misc
Misc01 <- subset(videogamesnewS, videogamesnewS$Genre == "Misc")
Misc <- sum (Misc01$Global_Sales)
Misc
```

```
## [1] 303.75
```

```r
#Género Racing
Racing01 <- subset(videogamesnewS, videogamesnewS$Genre == "Racing")
Racing <- sum (Racing01$Global_Sales)
Racing
```

```
## [1] 180.88
```

```r
#Género Fighting
Fighting01 <- subset(videogamesnewS, videogamesnewS$Genre == "Fighting")
Fighting <- sum (Fighting01$Global_Sales)
Fighting
```

```
## [1] 107.31
```

```r
#Género Platform
Platform01 <- subset(videogamesnewS, videogamesnewS$Genre == "Platform")
Platform <- sum (Platform01$Global_Sales)
Platform
```

```
## [1] 165.55
```

```r
#Género Puzzle
Puzzle01 <- subset(videogamesnewS, videogamesnewS$Genre == "Puzzle")
Puzzle <- sum (Puzzle01$Global_Sales)
Puzzle
```

```
## [1] 42.79
```

```
#Género Simulation
Simulation01 <- subset(videogamesnewS, videogamesnewS$Genre == "Simulation")
Simulation <- sum (Simulation01$Global_Sales)
Simulation
```

```
## [1] 115.37
```

```
#Género Strategy
Strategy01 <- subset(videogamesnewS, videogamesnewS$Genre == "Strategy")
Strategy <- sum (Strategy01$Global_Sales)
Strategy
```

```
## [1] 23.48
```

```
#Género Adventure
Adventure01 <- subset(videogamesnewS, videogamesnewS$Genre == "Adventure")
Adventure <- sum (Adventure01$Global_Sales)
Adventure
```

```
## [1] 49.11
```

```
#Análisis de los 5 géneros de videojuegos más vendidos
summary (c(Action,Sports,Shooter,RolePlaying,Misc))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   303.8   324.7   461.3   467.0   555.1   690.0
```

A continuación, vamos a revisar las ventas globales de acuerdo al editor de videojuego:

Ventas globales por Editor

```
#Editor Electronic Arts
ElectronicArts01 <- subset(videogamesnewS, videogamesnewS$Publisher == "Electronic Arts")
ElectronicArts <- sum (ElectronicArts01$Global_Sales)
ElectronicArts
```

```
## [1] 462.55
```

```
#Editor Nintendo
Nintendo01 <- subset(videogamesnewS, videogamesnewS$Publisher == "Nintendo")
Nintendo <- sum (Nintendo01$Global_Sales)
Nintendo
```

```
## [1] 529.24
```

```
#Editor Ubisoft
Ubisoft01 <- subset(videogamesnewS, videogamesnewS$Publisher == "Ubisoft")
Ubisoft <- sum (Ubisoft01$Global_Sales)
Ubisoft
```

```
## [1] 258.71
```

```
#Editor Activision
Activision01 <- subset(videogamesnewS, videogamesnewS$Publisher == "Activision")
Activision <- sum (Activision01$Global_Sales)
Activision
```

```
## [1] 378.94
```

```
#Editor Take-Two Interactive
TakeTwoInteractive01 <- subset(videogamesnewS,
```

```
                                videogamesnewS$Publisher == "Take-Two Interactive")
TakeTwoInteractive <- sum (TakeTwoInteractive01$Global_Sales)
TakeTwoInteractive
```

```
## [1] 209.61
```

```
#Editor Sony Computer Entertainment
SonyComputer01 <- subset(videogamesnewS,
                         videogamesnewS$Publisher == "Sony Computer Entertainment")
SonyComputer <- sum (SonyComputer01$Global_Sales)
SonyComputer
```

```
## [1] 176.61
```

```
#Editor THQ
THQ01 <- subset(videogamesnewS, videogamesnewS$Publisher == "THQ")
THQ <- sum (THQ01$Global_Sales)
THQ
```

```
## [1] 69.62
```

```
#Editor Sega
Sega01 <- subset(videogamesnewS, videogamesnewS$Publisher == "Sega")
Sega <- sum (Sega01$Global_Sales)
Sega
```

```
## [1] 87.16
```

```
#Editor Warner Bros
WarnerBros01 <- subset(videogamesnewS,
                       videogamesnewS$Publisher == "Warner Bros. Interactive Entertainment")
WarnerBros <- sum (WarnerBros01$Global_Sales)
WarnerBros
```

```
## [1] 93.18
```

```
#Análisis de los 5 editores con más videojuegos más vendidos
summary (c(ElectronicArts,Nintendo,Ubisoft,Activision,TakeTwoInteractive))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   209.6   258.7   378.9   367.8   462.6   529.2
```

En base a los datos observadores, se puede determinar que hay una empresa editora de videojuegos que lidera en los últimos 10 años, y las ventas de las otras 4 empresas considerablemente no se encuentran alejadas de la media de ventas.

Validación del género de los videojuegos del Top 50 de nuestro subconjunto de datos

```
videogamesrank<- subset(videogamesnewS, videogamesnewS$Rank < 51)
table(videogamesrank$Genre)
```

```
##
##       Action    Adventure     Fighting         Misc     Platform
##            4            0            1            1            2
##       Puzzle       Racing Role-Playing      Shooter   Simulation
##            0            2            3            9            0
##       Sports     Strategy
##            3            0
```

Validación de la plataforma de los videojuegos del Top 50 de nuestro subconjunto de datos

```
table(videogamesrank$Platform)
```

```
##
## 2600  3DO  3DS   DC   DS   GB  GBA   GC  GEN   GG  N64  NES   NG   PC PCFX
##    0    0    3    0    2    0    0    0    0    0    0    0    0    0    0
##   PS  PS2  PS3  PS4  PSP  PSV  SAT  SCD SNES TG16  Wii WiiU   WS X360   XB
##    0    0    4    2    0    0    0    0    0    0    7    0    0    7    0
## XOne
##    0
```

Adicional, vamos a realizar la prueba de Kruskal-Wallis de los valores globales con el resto de variables significativas:

```
#Género
kruskal.test (Genre ~ Global_Sales, data = videogames)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Genre by Global_Sales
## Kruskal-Wallis chi-squared = 656.28, df = 622, p-value = 0.1651
```

```
#Editor
kruskal.test (Publisher ~ Global_Sales, data = videogames)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Publisher by Global_Sales
## Kruskal-Wallis chi-squared = 590.83, df = 622, p-value = 0.8106
```

```
#Plataforma
kruskal.test (Platform ~ Global_Sales, data = videogames)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Platform by Global_Sales
## Kruskal-Wallis chi-squared = 669.77, df = 622, p-value = 0.09012
```

De acuerdo a los datos obtenidos, podemos deducir que al ser todos los p-valores > 0.05 se puede decir que los grupos son estadísticamente casi iguales, es decir no existen diferencias significativas entre las ventas globales y el resto de variables que estamos analizando del conjunto de datos.

4. Representación de los resultados a partir de tablas y gráficas.

A continuación, se presentarán los resultados obtenidos del análisis realizado a los registros de nuestro dataset limpio.

Plataforma de VideoJuegos

```
PlatformVideoGames <- cbind(PS3,X360,Wii,DS,PS4, EDS, XOne, PC, PS2)
colnames(PlatformVideoGames) <- c("PS3","X360","Wii","DS","PS4","3DS","XOne","PC","PS2")
kable(PlatformVideoGames)
```

PS3	X360	Wii	DS	PS4	3DS	XOne	PC	PS2

PS3	X360	Wii	DS	PS4	3DS	XOne	PC	PS2
699.74	711.19	553.91	294.39	232.28	175.37	105.76	83.43	75.56

Género de VideoJuegos

```
GenreVideoGames <- cbind(Action,Sports,Shooter,RolePlaying,Misc,Racing,Simulation,Fighting,
                         Platform,Puzzle)
colnames(GenreVideoGames) <- c("Action","Sports","Shooter","RolePlaying","Misc",
                               "Racing","Simulation","Fighting","Platform","Puzzle")
kable(GenreVideoGames)
```

Action	Sports	Shooter	RolePlaying	Misc	Racing	Simulation	Fighting	Platform	Puzzle
689.95	461.33	555.13	324.66	303.75	180.88	115.37	107.31	165.55	42.79

Editor de VideoJuegos

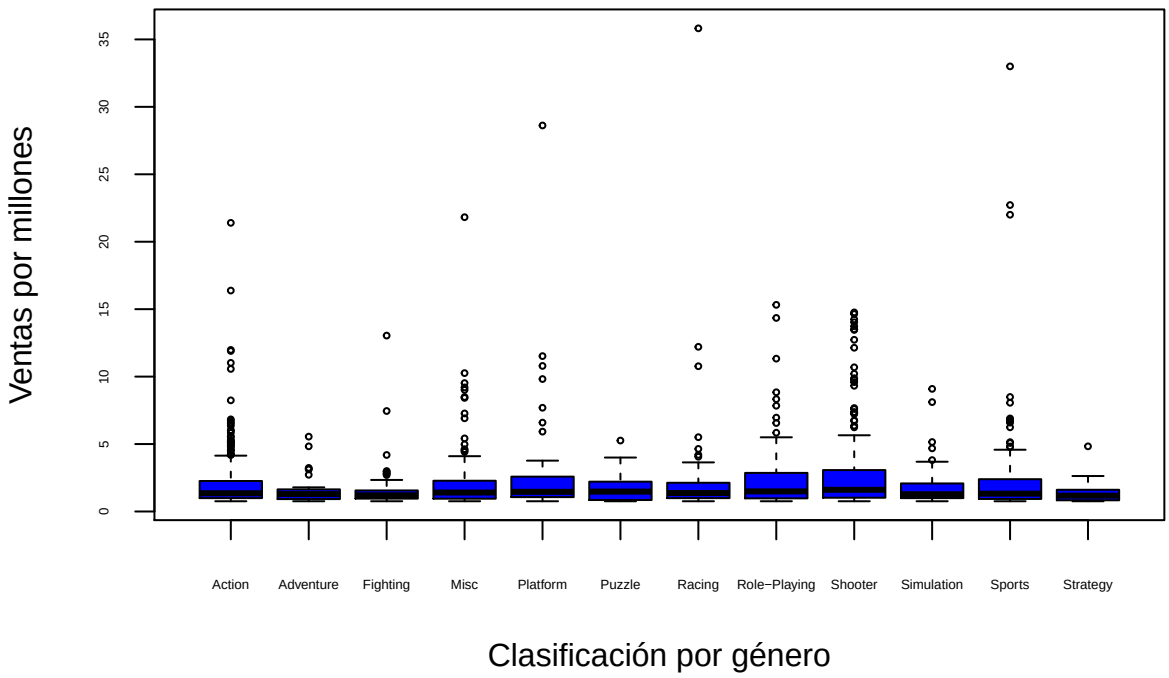
```
PublisherVideoGames <- cbind(ElectronicArts,Nintendo,Activision,Ubisoft,TakeTwoInteractive,
                         SonyComputer,WarnerBros,Sega)
colnames(PublisherVideoGames) <- c("ElectronicArts","Nintendo",
                                   "Activision","Ubisoft","Take2Interactive","SonyComputer",
                                   "WarnerBros","Sega")
kable(PublisherVideoGames)
```

ElectronicArts	Nintendo	Activision	Ubisoft	Take2Interactive	SonyComputer	WarnerBros	Sega
462.55	529.24	378.94	258.71	209.61	176.61	93.18	87.16

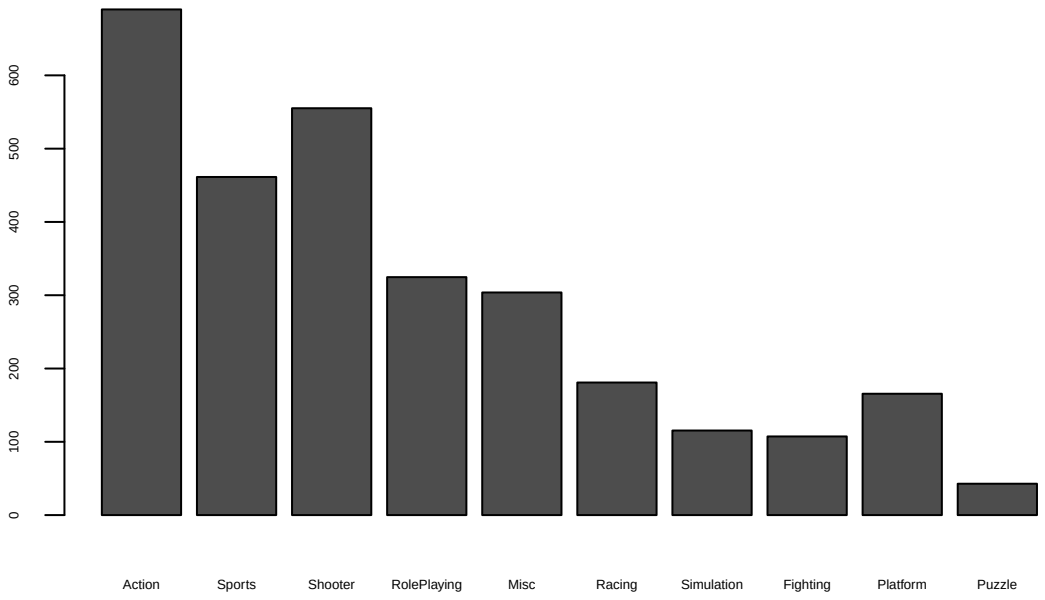
A continuación procedemos con la gráficación de las tres variables más influyentes del conjunto de datos en relación con las ventas globales.

Análisis de datos mediante gráficas

```
#Detalle de ventas por género de cada videojuego
plot(videogamesnewS$Genre, videogamesnewS$Global_Sales, col="blue",
     main = 'Diagrama de análisis de videojuegos por Género',
     ylab="Ventas por millones", xlab="Clasificación por género",cex.axis=0.4,cex=0.4)
```

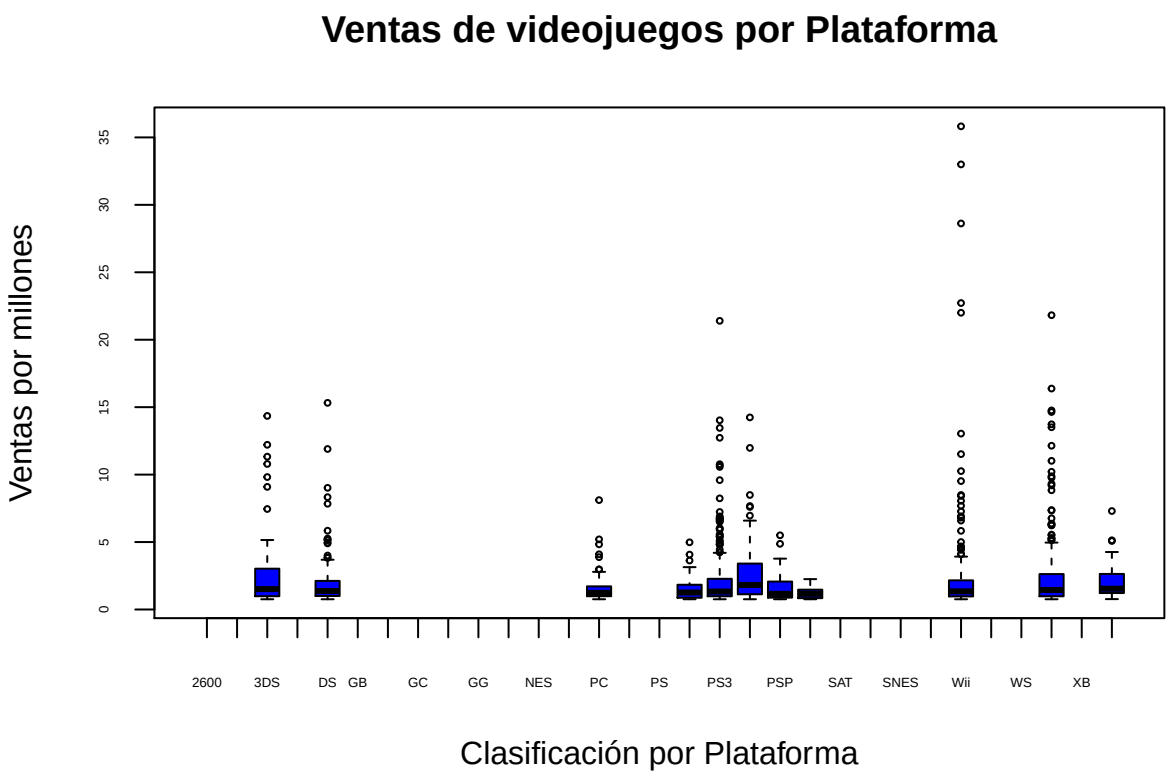



```
#Detalle global de ventas por cada género de videojuego
barplot(GenreVideoGames,cex.axis=0.4,cex=0.4)
```

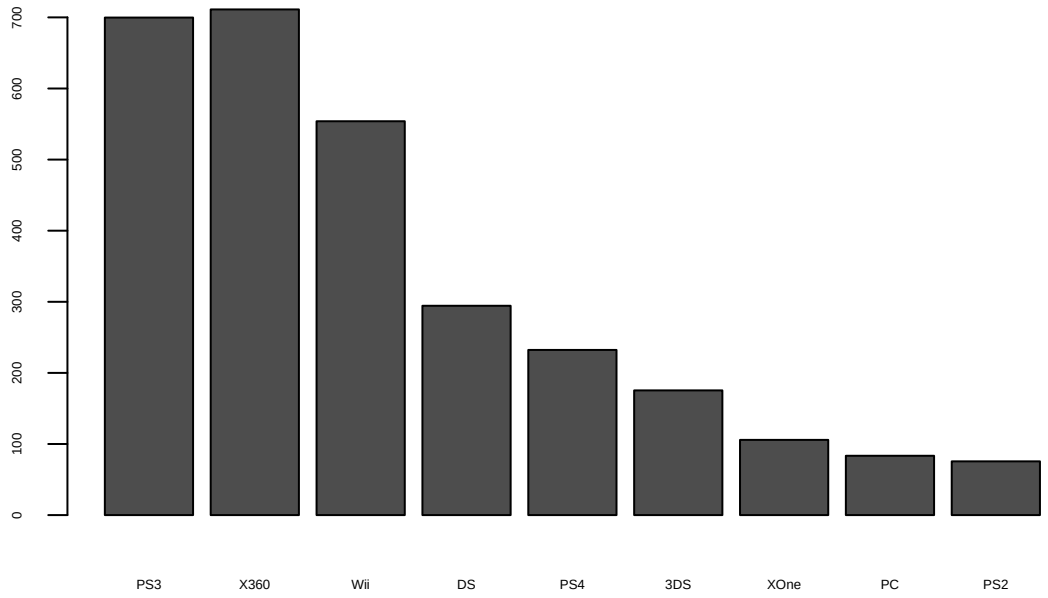


Conclusión: De acuerdo a la revisión de valores por género se puede decir que existen valores atípicos, pero al tratarse de ventas por unidad de videojuego única en el conjunto de datos se puede deducir aquellos géneros en videojuegos que triunfaron en el mercado, pero más allá no puede determinar el éxito absoluto de dicha categoría puesto que puede ser un "golpe de suerte" en la acogida que tuvo con el público. Entre los géneros que se observan que han tenido más perseverancia en el mercado son de acción, plataformas, disparos y deportes.

```
#Detalle de ventas por plataformas en videojuegos
plot(videogamesnewS$Platform, videogamesnewS$Global_Sales, col="blue",
     main = 'Ventas de videojuegos por Plataforma',
     ylab="Ventas por millones", xlab="Clasificación por Plataforma",cex.axis=0.4,cex=0.4)
```

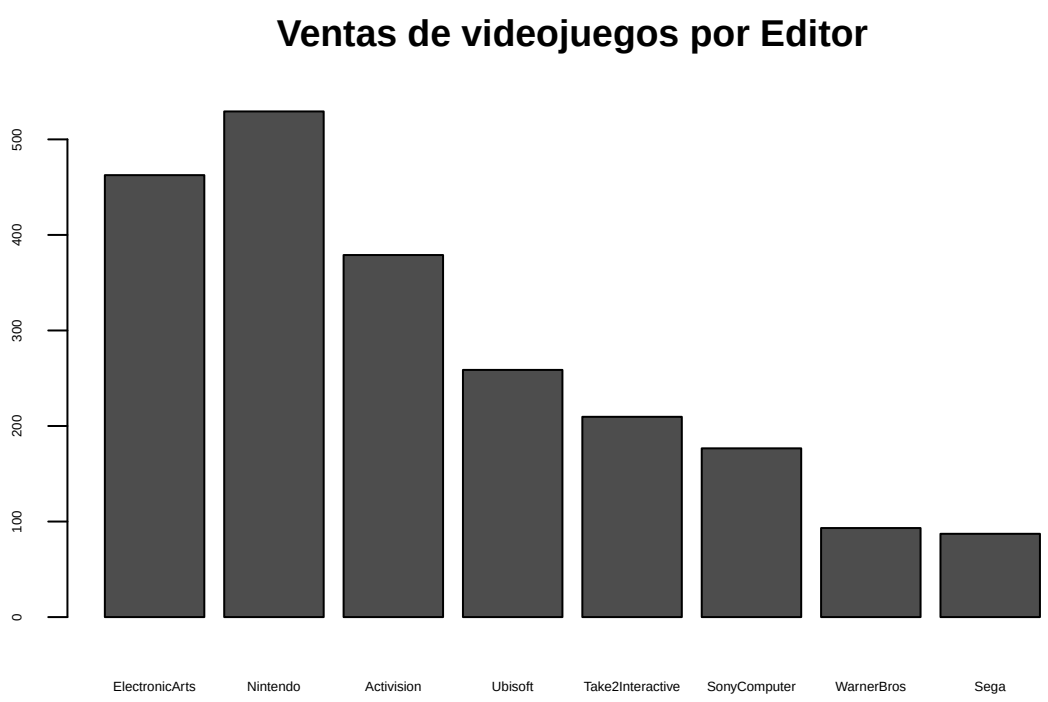



```
#Detalle global de ventas por plataforma
barplot(PlatformVideoGames,cex.axis=0.4,cex=0.4)
```



Conclusión: Al tratarse de plataformas de videojuegos, se conoce la evolución de las mismas y solo aquellas que innovan en su formato de presentación y adaptación de videojuegos para dicho tipo de consola sobreviven, en base a la gráfica se puede observar que existen plataformas con ventas bajas que se pueden considerar en la actualidad como obsoletas. Se puede evidenciar el éxito de determinadas plataformas entre ellas las versiones de Play Station y las versiones de Xbox, adicional se observa el éxito que tuvo la consola de Wii en su momento.

```
#Detalle global de Ventas por Editor
barplot(PublisherVideoGames, cex.axis=0.4,cex=0.4,main = 'Ventas de videojuegos por Editor')
```



Conclusión: De acuerdo a la revisión de ventas de editores, se evidencia aquellos con mayor éxito en el mercado, puede concluir que son aquellos que se mantienen vigentes de acuerdo al número de lanzamientos de videojuegos, lo que hay que considerar el número total de ventas puesto que pueden tener un par de juegos y figurarse entre los mejores.

De acuerdo a los datos graficados, se analizarán en conjuntos con los datos obtenidos de la revisión de los subconjuntos de datos por tipos que se reflejan en las tablas con la finalidad de fundamentar los resultados obtenidos y las propuestas realizadas.

Análisis por Plataforma y Género

Una vez analizados cada uno de los editores, géneros y plataformas por separado es necesario tener en cuenta que por cada plataforma existe una forma única de jugar, por lo cual es importante segmentar datos por grupos de plataformas similares para de esta forma, analizar sus ventas y géneros. En base a lo mencionado, se procede con la segmentación de tres grupos de datos:

- **Grupo 1:** PC, PS2, PS3, PS4, X360 Y XOne

```
Grupo01 <- subset(videogamesnewS, videogamesnewS$Platform == "PC" |
                    videogamesnewS$Platform == "PS2" | videogamesnewS$Platform == "PS3" |
                    videogamesnewS$Platform == "PS4" | videogamesnewS$Platform == "X360" |
                    videogamesnewS$Platform == "XOne")
table(Grupo01$Genre)
```

```
##
##       Action    Adventure     Fighting         Misc     Platform
##          242           12           53           59           23
##       Puzzle       Racing Role-Playing      Shooter   Simulation
##            0           56           74          171           17
```

```
##       Sports     Strategy
##          139           11
```

- **Grupo 2:** Wii

```
Grupo02 <- subset(videogamesnewS, videogamesnewS$Platform == "Wii")
table(Grupo02$Genre)
```

```
##
##       Action    Adventure     Fighting         Misc     Platform
##           36            6            5           51           18
##       Puzzle       Racing Role-Playing      Shooter   Simulation
##            5            8            3           11           13
##       Sports     Strategy
##           54            0
```

- **Grupo 3:** DS y 3DS

```
Grupo03 <- subset(videogamesnewS, videogamesnewS$Platform == "DS" |
                    videogamesnewS$Platform == "3DS")
table(Grupo03$Genre)
```

```
##
##       Action    Adventure     Fighting         Misc     Platform
##           46           11            2           21           15
##       Puzzle       Racing Role-Playing      Shooter   Simulation
##           19            4           47            3           30
##       Sports     Strategy
##            7            4
```

```
#Revisión de ventas significativas a partir de grupos de datos segmentados en plataformas
wilcox.test(Grupo01$Global_Sales,Grupo02$Global_Sales,paired=FALSE,conf.level=0.90)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Grupo01$Global_Sales and Grupo02$Global_Sales
## W = 94962, p-value = 0.2137
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(Grupo01$Global_Sales,Grupo03$Global_Sales,paired=FALSE,conf.level=0.90)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Grupo01$Global_Sales and Grupo03$Global_Sales
## W = 90808, p-value = 0.7538
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(Grupo02$Global_Sales,Grupo03$Global_Sales,paired=FALSE,conf.level=0.90)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Grupo02$Global_Sales and Grupo03$Global_Sales
## W = 21059, p-value = 0.4749
## alternative hypothesis: true location shift is not equal to 0
```

De acuerdo a los resultados de ventas analizados por grupos en plataformas, entre el grupo 1 y Grupo 2 no

existe mucha diferencia en cuanto al p-value, mientras que en la comparativa del grupo 1 y grupo 3 existe un valor-p elevado. En estas tres comparativas se evidencia que no cumple una hipótesis nula, lo que nos lleva a un resultado segmentado por plataforma.

5. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Con los datos obtenidos, es más fácil responder a una inquietud de una empresa en cuanto a la necesidad de crear un nuevo videojuego para posesionarse como una marca competente dentro del mercado actual.

En la limpieza de datos excluimos aquellos juegos con menos ganancias y aquellos videojuegos menores del año 2007 puesto que, 10 años después la forma de jugar ha cambiado y muchas de las consolas de videojuegos ya no se encuentran disponibles en el mercado. Además, los videojuegos se ven censurados de acuerdo a su clasificación de videojuegos puesto que actualmente están enfocados para ser utilizado por el público de todas las edades y el contenido debe ser restringido.

Bien, volviendo a nuestro propósito inicial hemos analizados cada uno de los datos de nuestro Conjunto de Datos, en donde llegamos a las siguientes conclusiones:

- **GÉNERO:** De acuerdo a los datos analizados en nuestro dataset, al final solo quedaron 12 categorías como las más demandadas en los últimos 10 años. De las cuales los más demandados figuran los géneros de acción, deportes, disparos (subcategoría de acción), Juego de roles y Misceláneas.

En la revisión de plataformas se identificó cambios en género de acuerdo al tipo de consola que se desea desarrollar, para lo cual se dividió las plataformas más demandadas en 3 grupos, esto de acuerdo al método de juegos y los accesorios que se disponen para manejar la plataforma que brindaN mejor experiencia al usuario, en donde se procede a recomendar lo siguiente:

- **Grupo 1 (PC, PS2, PS3, PS4, X360 Y XOne):** Acción, juegos de roles, carreras y peleas.
- **Grupo 2 (Wii):** Deportes, acción y estrategia.
- **Grupo 3 (DS y 3DS):** Acción, juego de roles y simulación.

Estos valores se encuentran fundamentados en ventas y aceptación de los usuarios. Hay que tener en cuenta, que para el desarrollo de cada grupo, existe la posibilidad de migrar las versiones a una nueva plafatorma disminuyendo así costos de desarrollo.

- **PLATAFORMA:** Para determinar a qué plataformas debemos enfocar el videojuego, se consideró un tiempo de 10 años que toma en dejar de comercializarse una consola, este también depende del nivel de acogida que tuvo por las personas, para lo cual al momento las plataformas más rentables son PS3, XBOX 360, y Wii. A pesar que, juegos para PS4 y XOne no son tan relevantes como las primeras es importante tomar en cuenta que son las sucesoras de las primeras consolas, y deben tomarse en cuenta para evitar un doble trabajo a futuro. Para desarrollar en plataformas DS y 3DS sería más recomendable de lanzar una versión móvil del juego. En este caso PC y PS2 ya no son tomadas en cuenta, puesto que un PC debe ser de buenas características para que se ejecute con normalidad un juego y PS2 ya salió del mercado. En base al número de ventas podría considerarse o no desarrollar para estas plataformas.
- **EDITOR:** En la parte del editor del videojuego, ese que será el encargado de plasmar nuestra idea en gráficos y código, la calidad del mismo dependerá mucho del capital que la empresa decida invertir. Para esto, una recomendación es trabajar en conjunto con Electronic Arts, Nintento, Activision y ubisoft. Adicional también estan Take Two interactive, Sony Computer, Warner Bros y Sega. Para fundamentarse en quien contratar la empresa puede basarse en los numeros de ventas de estas empresas.

Exportación de datos limpios

```
write.csv2(videogamesnewS$Platform , file="vgsales_clean.csv", row.names = FALSE)
```