Práctica 2: Limpieza y validación de los datos

Eduardo Ponce Palma

6 de enero de 2019

# Introduccion

## Presentacion

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas. Para hacer esta práctica tendréis que trabajar en grupos de hasta 3 personas, o si preferís, también podéis hacerlo de manera individual. Tendréis que entregar un solo archivo con el enlace Github (<https://github.com>) donde se encuentren las soluciones incluyendo los nombres de los componentes del equipo. Podéis utilizar la Wiki de Github para describir vuestro equipo y los diferentes archivos que corresponden a vuestra entrega. Cada miembro del equipo tendrá que contribuir con su usuario Github. Podéis utilizar estos ejemplos como guía:

● Ejemplo: <https://github.com/Bengis/nba-gap-cleaning>

● Ejemplo complejo (archivo adjunto).

## Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

● Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.

● Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

## Objetivos

Los objetivos concretos de esta práctica son:

● Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.

● Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.

● Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.

● Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.

● Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.

● Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.

● Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

# Practica

## Descripción

Para esta practica se ha tomado un conjunto de datos “Video Game Sales with Ratings” de la pagina web de KAGGLE.COM, este dataset fue generado hasta la fecha del 22 de diciembre del 2016 y representa un conjunto de datos de juegos, sus caracteristicas, sus ventas en el mundo y puntuaciones de Metacritic.

### El Dataset

El dataset se contituye de 16719 registros de 16 columna o atributos, se describen a continuacion:

Name.- Name of the game

Platform.- Console on which the game is running

Year\_of\_Release.- Año de lanzamiento del juego

Genre.- Categoria o genero del juego

Publisher.- Empresa que lo publica

NA\_Sales.- Ventas en Norte America (en millones/unidades)

EU\_Sales.- Ventas en la Union Europea (en millones/unidades)

JP\_Sales.- Ventas en Japon (en millones/unidades)

Other\_Sales.- Ventas en el resto de mundo excepto las anteriores (en millones/unidades)

Global\_Sales.- Total de ventas en el mundo (en millones/unidades)

Critic\_Score.- Puntuación agregada del personal de Metacritic

Critic\_Count.- Numero de criticas dependientes de Critic\_Score

User\_ScoreS.- Puntuación agregada de los suscriptores de Metacritic

User\_Count.- Numero de criticas dependientes de User\_ScoreS

Developer.- Desarrolladores del juego

Rating.- La calificacion del juego (adolecentes, solo adultos, etc)

### Importacia

En esta practica dado este conjunto de datos se pretende determinar que relacion tienen las ventas y las puntuaciones de Metacritic, se correlacionaran las variables para determinar si existe algun tipo de influencia.

Tambien se quiere determinar si los generos de mas popularidad en los ultimos 5 años tienen una diferencia de ventas significativa entre ellos.

## Integración y selección de los datos

En esta fase nos encargamos de extraer el dataset a ser examinado y cargamos los datos del archivo “VideoGames.csv”,

vg <- read.csv('VideoGames.csv', stringsAsFactors = FALSE)

Selecionamos los datos de interes y descartartamos las columnas que no necesitamos ya que no influyen para el proposito del analisis, luego verificamos la estructura del juego de datos.

vg$Name <- NULL  
vg$Publisher <- NULL  
vg$Developer <- NULL  
vg$Rating <- NULL  
str(vg)

## 'data.frame': 16719 obs. of 12 variables:  
## $ Platform : chr "Wii" "NES" "Wii" "Wii" ...  
## $ Year\_of\_Release: chr "2006" "1985" "2008" "2009" ...  
## $ Genre : chr "Sports" "Platform" "Racing" "Sports" ...  
## $ 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 : chr "8" "" "8.3" "8" ...  
## $ User\_Count : int 322 NA 709 192 NA NA 431 129 594 NA ...

## Limpieza de los datos

En esta fase se tratara de resolver los datos no consistentes o datos no validos, primero se verificara que columnas los contienen.

colSums(is.na(vg))

## Platform Year\_of\_Release Genre NA\_Sales   
## 0 0 0 0   
## EU\_Sales JP\_Sales Other\_Sales Global\_Sales   
## 0 0 0 0   
## Critic\_Score Critic\_Count User\_Score User\_Count   
## 8582 8582 0 9129

colSums(vg=="N/A")

## Platform Year\_of\_Release Genre NA\_Sales   
## 0 269 0 0   
## EU\_Sales JP\_Sales Other\_Sales Global\_Sales   
## 0 0 0 0   
## Critic\_Score Critic\_Count User\_Score User\_Count   
## NA NA 0 NA

colSums(vg=="0")

## Platform Year\_of\_Release Genre NA\_Sales   
## 0 0 0 4511   
## EU\_Sales JP\_Sales Other\_Sales Global\_Sales   
## 5874 10515 6604 0   
## Critic\_Score Critic\_Count User\_Score User\_Count   
## NA NA 1 NA

colSums(vg=="")

## Platform Year\_of\_Release Genre NA\_Sales   
## 0 0 2 0   
## EU\_Sales JP\_Sales Other\_Sales Global\_Sales   
## 0 0 0 0   
## Critic\_Score Critic\_Count User\_Score User\_Count   
## NA NA 6704 NA

Una vez realizada la inspeccion procedemos a la limpieza de datos, procedemos a eliminar los valores extremos de Year\_of\_Release mayores al 2016 ya que el dataset se formo el 22 de diciembre del 2016 y cambianos los valores de “N/A” a nulos para luego hacer el conjunto a entero y colocar valores de la mediana en los valores nulos, los Géneros vacios se colocan el nombre de “S/G” (sin genero) y finalmente cambianos los valores de “tbd” de User\_Score a nulos para hacerla conjunto entero.

Los demas valores extremos identificados corresponden a valores de sumatorias de ventas o numeros de criticas y por ende son valores reales posibles.

Mostraremos para cada atributo los valores mediante la funcion summary.

vg <- vg[vg$Year\_of\_Release<=2016,]  
vg$Year\_of\_Release[vg$Year\_of\_Release=="N/A"]=NA  
vg$Year\_of\_Release <- as.integer(vg$Year\_of\_Release)  
vg$Year\_of\_Release[is.na(vg$Year\_of\_Release)] <- median(vg$Year\_of\_Release,na.rm=T)  
  
vg$Genre[vg$Genre==""]="S/G"  
  
vg$User\_Score[vg$User\_Score=="tbd"]=NA  
vg$User\_Score <- as.integer(vg$User\_Score)  
  
summary(vg)

## Platform Year\_of\_Release Genre NA\_Sales   
## Length:16446 Min. :1980 Length:16446 Min. : 0.0000   
## Class :character 1st Qu.:2003 Class :character 1st Qu.: 0.0000   
## Mode :character Median :2007 Mode :character Median : 0.0800   
## Mean :2006 Mean : 0.2641   
## 3rd Qu.:2010 3rd Qu.: 0.2400   
## Max. :2016 Max. :41.3600   
##   
## EU\_Sales JP\_Sales Other\_Sales Global\_Sales   
## Min. : 0.0000 Min. : 0.00000 Min. : 0.00000 Min. : 0.0100   
## 1st Qu.: 0.0000 1st Qu.: 0.00000 1st Qu.: 0.00000 1st Qu.: 0.0600   
## Median : 0.0200 Median : 0.00000 Median : 0.01000 Median : 0.1700   
## Mean : 0.1459 Mean : 0.07848 Mean : 0.04759 Mean : 0.5364   
## 3rd Qu.: 0.1100 3rd Qu.: 0.04000 3rd Qu.: 0.03000 3rd Qu.: 0.4700   
## Max. :28.9600 Max. :10.22000 Max. :10.57000 Max. :82.5300   
##   
## Critic\_Score Critic\_Count User\_Score User\_Count   
## Min. :13.00 Min. : 3.00 Min. :0.00 Min. : 4   
## 1st Qu.:60.00 1st Qu.: 12.00 1st Qu.:6.00 1st Qu.: 10   
## Median :71.00 Median : 22.00 Median :7.00 Median : 24   
## Mean :68.99 Mean : 26.44 Mean :6.68 Mean : 163   
## 3rd Qu.:79.00 3rd Qu.: 36.00 3rd Qu.:8.00 3rd Qu.: 81   
## Max. :98.00 Max. :113.00 Max. :9.00 Max. :10665   
## NA's :8463 NA's :8463 NA's :8983 NA's :8983

Adicionamente exportamos el dataset generado despues de la limpieza de datos

write.csv(vg, "VideoGamesClean.csv")

## Análisis de los datos

En esta fase se realizaran los proceso de analisis para poder responder a las preguntas planteadas en el inicio de la practica.

### Selección de grupos

En esta fase selecionamos a los juegos publicados desde el 2011, para poder observar la tendencia de los ultimos años de desarrollo de juegos respecto al genero, no se tomara en cuenta la plataforma puesto que es cambiante y cada cierto tiempo hay nuevas consolas.

vg\_5y <- vg[vg$Year\_of\_Release>2010,]  
summary(vg\_5y)

## Platform Year\_of\_Release Genre NA\_Sales   
## Length:4022 Min. :2011 Length:4022 Min. :0.0000   
## Class :character 1st Qu.:2011 Class :character 1st Qu.:0.0000   
## Mode :character Median :2013 Mode :character Median :0.0400   
## Mean :2013 Mean :0.2063   
## 3rd Qu.:2015 3rd Qu.:0.1700   
## Max. :2016 Max. :9.6600   
##   
## EU\_Sales JP\_Sales Other\_Sales Global\_Sales   
## Min. :0.0000 Min. :0.00000 Min. :0.00000 Min. : 0.010   
## 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.: 0.040   
## Median :0.0300 Median :0.00000 Median :0.01000 Median : 0.130   
## Mean :0.1665 Mean :0.06108 Mean :0.05194 Mean : 0.486   
## 3rd Qu.:0.1300 3rd Qu.:0.05000 3rd Qu.:0.04000 3rd Qu.: 0.400   
## Max. :9.0900 Max. :4.39000 Max. :3.96000 Max. :21.040   
##   
## Critic\_Score Critic\_Count User\_Score User\_Count   
## Min. :13.00 Min. : 3.00 Min. :0.000 Min. : 4.0   
## 1st Qu.:64.00 1st Qu.: 13.00 1st Qu.:5.000 1st Qu.: 23.0   
## Median :74.00 Median : 27.00 Median :7.000 Median : 74.0   
## Mean :71.27 Mean : 31.72 Mean :6.191 Mean : 347.4   
## 3rd Qu.:81.00 3rd Qu.: 46.00 3rd Qu.:7.000 3rd Qu.: 278.2   
## Max. :97.00 Max. :113.00 Max. :9.000 Max. :10665.0   
## NA's :2210 NA's :2210 NA's :1990 NA's :1990

str(vg\_5y)

## 'data.frame': 4022 obs. of 12 variables:  
## $ Platform : chr "PS3" "X360" "X360" "PS4" ...  
## $ Year\_of\_Release: num 2013 2013 2011 2015 2013 ...  
## $ Genre : chr "Action" "Action" "Shooter" "Shooter" ...  
## $ NA\_Sales : num 7.02 9.66 9.04 6.03 5.28 4.99 8.25 5.54 5.03 3.96 ...  
## $ EU\_Sales : num 9.09 5.14 4.24 5.86 4.19 5.73 4.24 5.73 4.02 6.31 ...  
## $ JP\_Sales : num 0.98 0.06 0.13 0.36 4.35 0.65 0.07 0.49 2.69 0.38 ...  
## $ Other\_Sales : num 3.96 1.41 1.32 2.38 0.78 2.42 1.12 1.57 0.91 1.97 ...  
## $ Global\_Sales : num 21 16.3 14.7 14.6 14.6 ...  
## $ Critic\_Score : int 97 97 88 NA NA 83 83 88 85 97 ...  
## $ Critic\_Count : int 50 58 81 NA NA 21 73 39 73 66 ...  
## $ User\_Score : int 8 8 3 NA NA 5 4 3 8 8 ...  
## $ User\_Count : int 3994 3711 8713 NA NA 922 2256 5234 632 2899 ...

Selecionamos los 3 primeros generos que son “Action”, “Role-Playing” y “Adventure” cada uno en un subconjunto.

vg\_Genero1 <- vg\_5y[vg\_5y$Genre == "Action",]$Global\_Sales  
vg\_Genero2 <- vg\_5y[vg\_5y$Genre == "Role-Playing",]$Global\_Sales  
vg\_Genero3 <- vg\_5y[vg\_5y$Genre == "Adventure",]$Global\_Sales

### Comprobación de la normalidad y homogeneidad de la varianza

Se comprueba si existe atributos con distribucion normal y utilizamos la prueba de AndersonDarling ya que la muestra es mayor a 5000 registros. Se fija α = 0.05, este algoritmo fue tomado del ejemplo de la practica.

library(nortest)

## Warning: package 'nortest' was built under R version 3.5.2

alpha = 0.05  
col.names = colnames(vg)  
for (i in 1:ncol(vg)) {  
 if (i == 1) cat("Variables que no siguen una distribución normal:\n")  
   
 if (is.integer(vg[,i]) | is.numeric(vg[,i])) {  
 p\_val = ad.test(vg[,i])$p.value  
 if (p\_val < alpha) {  
 cat(col.names[i])  
 # Format output  
 if (i < ncol(vg) - 1) cat(", ")  
 if (i %% 3 == 0) cat("\n")  
 }  
 }  
}

## Variables que no siguen una distribución normal:  
## Year\_of\_Release, NA\_Sales, EU\_Sales, JP\_Sales,   
## Other\_Sales, Global\_Sales, Critic\_Score,   
## Critic\_Count, User\_ScoreUser\_Count

En este caso ningun atributo siguió la distribucion normal y por esta razón no se puede realizar la homogeneidad de la varianza.

Aqui una prueba de que el valor de “p-value” tiene de a cero

vg$Platform <- as.factor(vg$Platform)  
fligner.test(Year\_of\_Release ~ Platform, data = vg)

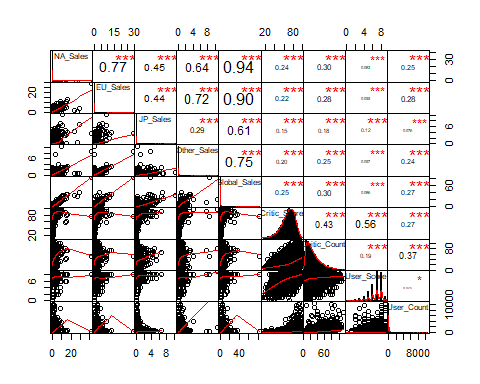
##   
## Fligner-Killeen test of homogeneity of variances  
##   
## data: Year\_of\_Release by Platform  
## Fligner-Killeen:med chi-squared = 2841.9, df = 30, p-value <  
## 2.2e-16

Se muestra que el valor de p-value es menor a 0.05

### Aplicación de pruebas estadísticas

En esta fase creamos una matriz de correlacion entre la variables numericas para determinar si existe una relacion entre la muestras,

library(PerformanceAnalytics)  
vg\_grupo <- vg[,4:12]  
#vg\_grupo <- na.omit(vg\_grupo)  
chart.Correlation(vg\_grupo)



Aplicado la matriz podemos decir que existe una fuerte relacion entre la venta generadas en NA\_Sales y EU\_Sales contra Global\_Sales, que las ventas tiene poca relacion con la puntuaciones de Critic\_Score y casi nula con las puntuaciones de User\_Score

Tambien se realizan pruebas de comparación entre los 3 primeros generos de mayor popularidad desde el 2011.

t.test(vg\_Genero1, vg\_Genero2,alternative = "less")

##   
## Welch Two Sample t-test  
##   
## data: vg\_Genero1 and vg\_Genero2  
## t = -1.2975, df = 731.86, p-value = 0.09743  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
## -Inf 0.02378093  
## sample estimates:  
## mean of x mean of y   
## 0.4397163 0.5280215

t.test(vg\_Genero1, vg\_Genero3,alternative = "less")

##   
## Welch Two Sample t-test  
##   
## data: vg\_Genero1 and vg\_Genero3  
## t = 9.588, df = 1616.3, p-value = 1  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
## -Inf 0.3863987  
## sample estimates:  
## mean of x mean of y   
## 0.4397163 0.1099268

t.test(vg\_Genero2, vg\_Genero3,alternative = "less")

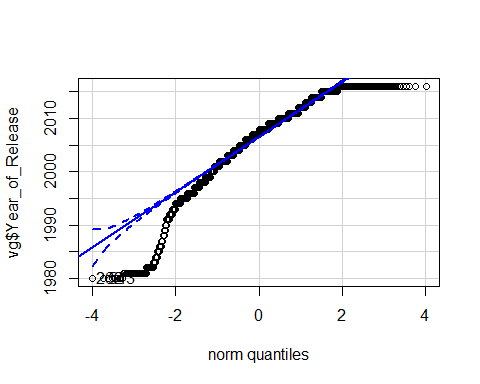
##   
## Welch Two Sample t-test  
##   
## data: vg\_Genero2 and vg\_Genero3  
## t = 6.7568, df = 512.16, p-value = 1  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
## -Inf 0.5200591  
## sample estimates:  
## mean of x mean of y   
## 0.5280215 0.1099268

Comparados los resultados se puede observar que las ventas globales no dependen de genero ya que sus “p-value” es mayor a 0.05, esto quiere decir que los principales generos de los ultimos 5 años no generan mas ventas que los demas.

## resultados

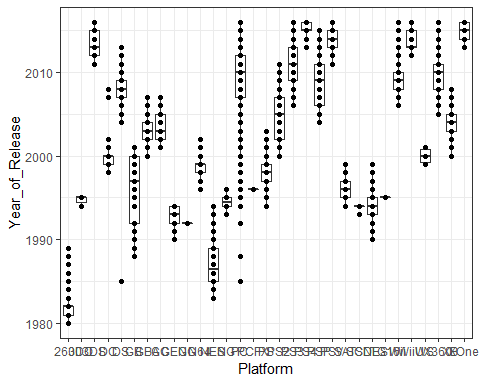
Mostramos los resultados del atributo “Year\_of\_Release” para ver graficamente que no existe distribucion normal y que la homogeneidad de la varianza del mismo respecto a la plataforma no se da el resultado.

# distribucion no normal  
library("car")  
library(ggplot2)  
qqPlot(vg$Year\_of\_Release)



## [1] 262 543

# homogeneidad de la varianza  
ggplot(data = vg, aes(x = Platform, y = Year\_of\_Release)) + geom\_boxplot() + geom\_point() + theme\_bw() + theme(legend.position = "none")



## Concluciones

Como podemos observar en la practica, donde se ha podido analizar el conjunto de datos y se ha podido determinar que:

1.- No existe relacion significativa entre la ventas de los juegos y las puntuaciones de Metacritic, lo que quiere decir que una mala o buena critica no influira en mayor medida con las ventas en cualquier parte del mundo, pero se puedo descubrir que si existe una relacion entre las ventas de norte america o la union europea con las ventas globales y se podria concluir que si se logra buenas ventas en estas zonas tendrá influencia positiva en el resto del mundo.

2.- Tambien se pudo determinar que los generos de más popularidad en los ultimos 5 años no generan ma ventas globales que otros, asi que es posible ir por el desarrollo o compra de juegos de cualquier genero y su rentabilidad puede ser la misma, eso dependera posiblemente de otros factores que no se han proporcionado en este conjunto de datos.