AnalisExploratorioVideoGames

Victor René Cortez Sis & Gerson Omar Pérez

24/5/2021

## Enlace del Data set: <https://www.kaggle.com/sidtwr/videogames-sales-dataset?select=XboxOne_GameSales.csv>

## Librerias

require(stats)  
library(dplyr)

## Warning: package 'dplyr' was built under R version 4.0.5

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.0.5

library(readr)

## Warning: package 'readr' was built under R version 4.0.5

library(nortest) # prueba de normalidad de Anderson-Darling, que funciona para tamaños de muestra más grandes

## Cargar y leer el archivo

VideoGames<-read.csv("Video\_Games\_Sales\_as\_at\_22\_Dec\_2016.csv")

## Explorar la estructura y las variables

str(VideoGames)

## 'data.frame': 16719 obs. of 16 variables:  
## $ Name : chr "Wii Sports" "Super Mario Bros." "Mario Kart Wii" "Wii Sports Resort" ...  
## $ Platform : chr "Wii" "NES" "Wii" "Wii" ...  
## $ Year\_of\_Release: chr "2006" "1985" "2008" "2009" ...  
## $ Genre : chr "Sports" "Platform" "Racing" "Sports" ...  
## $ Publisher : chr "Nintendo" "Nintendo" "Nintendo" "Nintendo" ...  
## $ 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 ...  
## $ Developer : chr "Nintendo" "" "Nintendo" "Nintendo" ...  
## $ Rating : chr "E" "" "E" "E" ...

## Ver las primeras filas de los datos

head(VideoGames)

## Name Platform Year\_of\_Release Genre Publisher  
## 1 Wii Sports Wii 2006 Sports Nintendo  
## 2 Super Mario Bros. NES 1985 Platform Nintendo  
## 3 Mario Kart Wii Wii 2008 Racing Nintendo  
## 4 Wii Sports Resort Wii 2009 Sports Nintendo  
## 5 Pokemon Red/Pokemon Blue GB 1996 Role-Playing Nintendo  
## 6 Tetris GB 1989 Puzzle Nintendo  
## NA\_Sales EU\_Sales JP\_Sales Other\_Sales Global\_Sales Critic\_Score Critic\_Count  
## 1 41.36 28.96 3.77 8.45 82.53 76 51  
## 2 29.08 3.58 6.81 0.77 40.24 NA NA  
## 3 15.68 12.76 3.79 3.29 35.52 82 73  
## 4 15.61 10.93 3.28 2.95 32.77 80 73  
## 5 11.27 8.89 10.22 1.00 31.37 NA NA  
## 6 23.20 2.26 4.22 0.58 30.26 NA NA  
## User\_Score User\_Count Developer Rating  
## 1 8 322 Nintendo E  
## 2 NA   
## 3 8.3 709 Nintendo E  
## 4 8 192 Nintendo E  
## 5 NA   
## 6 NA

## Obtener un resumen estadistico de cada variables

summary(VideoGames)

## Name Platform Year\_of\_Release Genre   
## Length:16719 Length:16719 Length:16719 Length:16719   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## Publisher NA\_Sales EU\_Sales JP\_Sales   
## Length:16719 Min. : 0.0000 Min. : 0.000 Min. : 0.0000   
## Class :character 1st Qu.: 0.0000 1st Qu.: 0.000 1st Qu.: 0.0000   
## Mode :character Median : 0.0800 Median : 0.020 Median : 0.0000   
## Mean : 0.2633 Mean : 0.145 Mean : 0.0776   
## 3rd Qu.: 0.2400 3rd Qu.: 0.110 3rd Qu.: 0.0400   
## Max. :41.3600 Max. :28.960 Max. :10.2200   
##   
## Other\_Sales Global\_Sales Critic\_Score Critic\_Count   
## Min. : 0.00000 Min. : 0.0100 Min. :13.00 Min. : 3.00   
## 1st Qu.: 0.00000 1st Qu.: 0.0600 1st Qu.:60.00 1st Qu.: 12.00   
## Median : 0.01000 Median : 0.1700 Median :71.00 Median : 21.00   
## Mean : 0.04733 Mean : 0.5335 Mean :68.97 Mean : 26.36   
## 3rd Qu.: 0.03000 3rd Qu.: 0.4700 3rd Qu.:79.00 3rd Qu.: 36.00   
## Max. :10.57000 Max. :82.5300 Max. :98.00 Max. :113.00   
## NA's :8582 NA's :8582   
## User\_Score User\_Count Developer Rating   
## Length:16719 Min. : 4.0 Length:16719 Length:16719   
## Class :character 1st Qu.: 10.0 Class :character Class :character   
## Mode :character Median : 24.0 Mode :character Mode :character   
## Mean : 162.2   
## 3rd Qu.: 81.0   
## Max. :10665.0   
## NA's :9129

## Pregunta 1:

## ¿Sera que las ventas en Norteámerica van a depender al tipo de genero de video juegos. Es decir que tipo de genero de video juegos venden mas en Norteámerica?

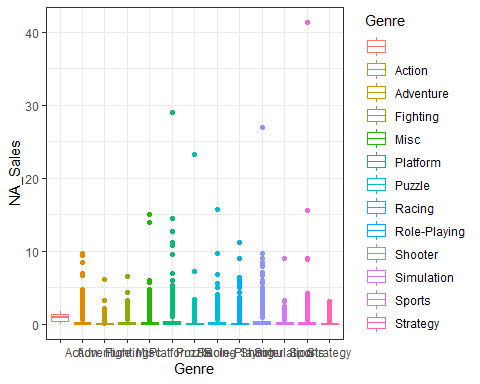
## NA\_Sales= Cuantitativa

## genre= Cualitativa

## Graficamente = Boxplot

## Modelo: Anova

ggplot(data=VideoGames, aes(x=Genre, y=NA\_Sales, color=Genre))+geom\_boxplot()+theme\_bw()



## Como se observa en la grafica vemos el comportamiento de las tipos de generos en funciones a la ventas globales lo cual algunos generos tiene una distribucion diferente de cada por que prodia ser el genero de videosjuegos mas vendiso es en plataforma y esta el otro genero tambien que es sports.

## Hipotesis

## No hay diferencia entre las medias del los grupos

## Al menor un par de medias son significativas distintas la una de la otra

anova<-aov(VideoGames$NA\_Sales~VideoGames$Genre)

summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## VideoGames$Genre 12 169 14.077 21.59 <2e-16 \*\*\*  
## Residuals 16706 10895 0.652   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Pr(>F)=2e-16 \*\*\*

## Respuesta: Dado que el p-value (2e-16 < 0.05) es mayor al nivel de significancia, no hay suficiente evidenica para afirmar que al menos dos medias son distintas entre genero de videojuegos.

## Pregunta 2:

## ¿Sera que las ventas en Europa van a depender al tipo de genero de video juegos. Es decir que tipo de genero de video juegos venden mas en Europa?

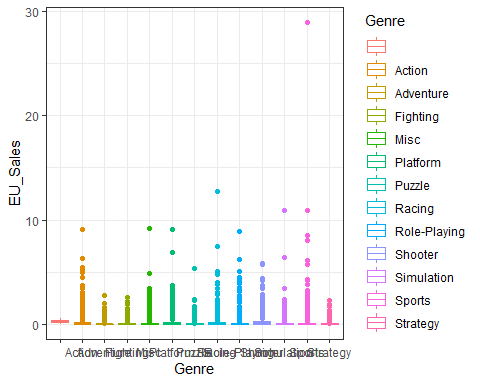
## EU\_Sales= Cuantitativa

## genre= Cualitativa

## Graficamente = Boxplot

## Modelo: Anova

ggplot(data=VideoGames, aes(x=Genre, y=EU\_Sales, color=Genre))+geom\_boxplot()+theme\_bw()



## Como se observa en la grafica vemos los comportamientos de los generos de los videojuegos. Que los generos varian tambien al precio de ventas en europa.

## Hipotesis

## No hay diferencia entre las medias del los grupos

## Al menor un par de medias son significativas distintas la una de la otra

anova<-aov(VideoGames$EU\_Sales~VideoGames$Genre)

summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## VideoGames$Genre 12 42 3.470 13.83 <2e-16 \*\*\*  
## Residuals 16706 4193 0.251   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Respesta: Como nuestro p-values (2e-16 < 0.05), no se rechza la hipotesis nula y no hay suficiente evidencua para afirmar que la menos un par de medias son distintas entre el genero.

## Pregunta 3:

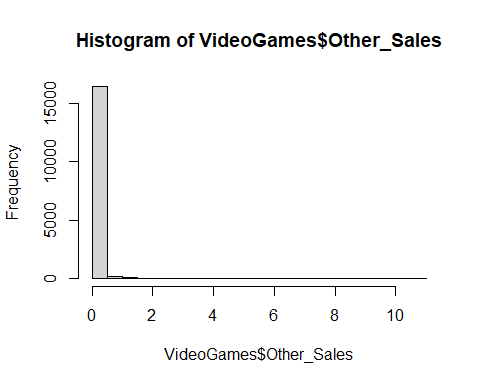
## ¡Explorar la variable OtrasVentas(Other\_Sales) que se refiere a las ventas de video juegos en el resto del mundo (en millones ) para para determinar la normalidad de los datos?

## Other\_Sales = Cuantitativa

## Graficamente = Histograma

## MOdelo = Shapiro

hist(VideoGames$Other\_Sales)



## En la grafica se ve claramente que tiene un distribucion sesgada a la izquiereda.

## Hipotesis

## Ho: Normalidad de los datos

## H1: No hay normalidad de los datos

ad.test(VideoGames$Other\_Sales)

##   
## Anderson-Darling normality test  
##   
## data: VideoGames$Other\_Sales  
## A = 3860.6, p-value < 2.2e-16

2.2e-16<0.05

## [1] TRUE

## Respuesta: Como el p-value es mayo que el nivel de significancia 0.05, no se recha la hipotesis Ho, por ende hay normalidad en los datos.

## Pregunta 4:

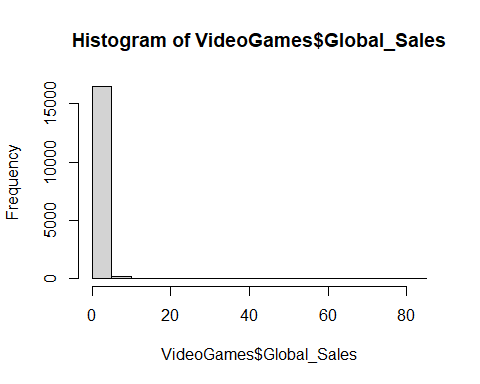
## Explorar de la variable VentasGlobales(Global\_Sales) que se refiere a las ventas de video juegos mundiales totales para determinar la normalidad de los datos.

## Global\_Sales = Cuantitativa

## Graficamente = Histograma

## MOdelo = Shapiro

hist(VideoGames$Global\_Sales)



## Vemos el la grafica que no tiene un distibucion normal porque esta sesgada a la izquierda, vemos la frecuncia de frecuencia de las ventas que por lo menos entre 0 a 5 vemos que la venta en europa de videojuegos es bastante alta.

## Hipotesis

## Ho: Normalidad de los datos

## H1: No hay normalidad de los datos

ad.test(VideoGames$Global\_Sales)

##   
## Anderson-Darling normality test  
##   
## data: VideoGames$Global\_Sales  
## A = 3313.4, p-value < 2.2e-16

## Respuesta: dado que el p-value es yor que el nivel de significancia, no se no rechaza la hipoteis Ho, por ende hay normalidad en los datos de las ventas globales.

## Pregunta 5:

## ¿Sera que las ventas en Jápon van a depender al tipo de genero de video juegos. Es decir que tipo de genero de video juegos venden mas en Jápon?

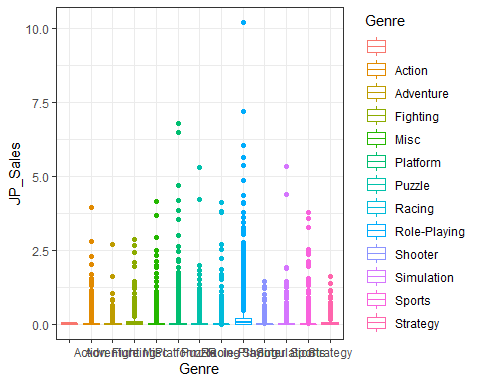
## JP\_Sales= Cuantitativa

## Genre= Cualitativa

## Graficamente = Boxplot

## Modelo: Anova

ggplot(data = VideoGames, aes(x = Genre, y=JP\_Sales, color=Genre))+ geom\_boxplot() + theme\_bw()



## Como se observa que los datos de tipos de genero tienen un comportamiento diferente por lo que dependen de las ventas, pero mas sin embargo el tamaño de las cajas nos indica variablidad en los datos entre genero de videojuegos.

## Hipotesis:

## No hay diferencia entre las medias del los grupos

## Al menor un par de medias son significativas distintas la una de la otra

anova<-aov(VideoGames$JP\_Sales~VideoGames$Genre)

anova

## Call:  
## aov(formula = VideoGames$JP\_Sales ~ VideoGames$Genre)  
##   
## Terms:  
## VideoGames$Genre Residuals  
## Sum of Squares 53.8179 1540.5544  
## Deg. of Freedom 12 16706  
##   
## Residual standard error: 0.3036703  
## Estimated effects may be unbalanced

summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## VideoGames$Genre 12 53.8 4.485 48.63 <2e-16 \*\*\*  
## Residuals 16706 1540.6 0.092   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Respuesta: Debido a que nuestro p-values es mayor al nivel de significancia, no se rechaz la hipotesis Ho, por lo que no hay suficiente evidencia para un par de las medias son significativas distrintas entre genero.

## Pregunta 6:

## Realizar un analisis explorario de la variable clasificacion (Rating) en función a los tipos de genero del juego. Para determinar la clasificaciones del tipo de genero de juegos es decir clasificarlos en categoria para menores de edad el juego,el contenido por lo genera es apto para todas las edades, clasificar los videosjuegos segun su contenido de violencia fisica o verbal, entre otras clasificaciones.

## Rating = Cuatitativa

## Genre = Cuatitativa

## Graficamente = Mosaico

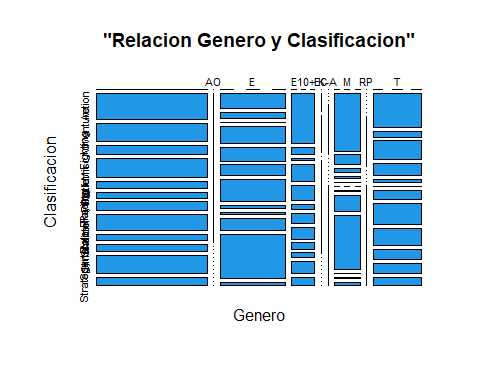
## Modelo = chi cuadrado

tabla<-table(VideoGames$Rating, VideoGames$Genre, dnn = c("Genero", "Clasificacion"))

tabla

## Clasificacion  
## Genero Action Adventure Fighting Misc Platform Puzzle Racing Role-Playing  
## 2 1182 857 411 868 319 238 377 723  
## AO 0 1 0 0 0 0 0 0 0  
## E 0 416 162 8 457 358 289 585 84  
## E10+ 0 481 68 19 167 144 43 96 111  
## EC 0 1 2 0 5 0 0 0 0  
## K-A 0 0 0 0 1 0 0 0 0  
## M 0 608 99 49 13 3 0 18 162  
## RP 0 0 0 0 0 0 0 1 0  
## T 0 681 115 362 239 64 10 172 420  
## Clasificacion  
## Genero Shooter Simulation Sports Strategy  
## 304 305 839 344  
## AO 0 0 0 0  
## E 48 326 1188 70  
## E10+ 58 48 107 78  
## EC 0 0 0 0  
## K-A 0 0 0 2  
## M 565 5 16 25  
## RP 0 0 0 2  
## T 348 190 198 162

mosaicplot(tabla, color=4.2, main = deparse("Relacion Genero y Clasificacion"))



## Este grafico no representa la clasificacion de los tipos de generos.

## Ho: Son variables independientes (No hay relacion)

## H1: No son variables independientes (Si hay relacion)

chisq.test(tabla)

## Warning in chisq.test(tabla): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: tabla  
## X-squared = 6487.3, df = 96, p-value < 2.2e-16

# Respuesta: como nuestro p-value es mayor al nivel de signicancia, no se rechaza la hipotesis Ho, por lo que no hay evidencia suficiente para confirmar que exista una relacion entre clasificacion y el genero de juegos.

## Pregunta 7:

## ¿sera que el tipo de plataforma de lanzamineto del juego (Platform) tiene relacion con las ventas en Norteámerica (en millones).Es decir cual es el tipo de plataforma tiene que ver con las ventas en norteamerica?

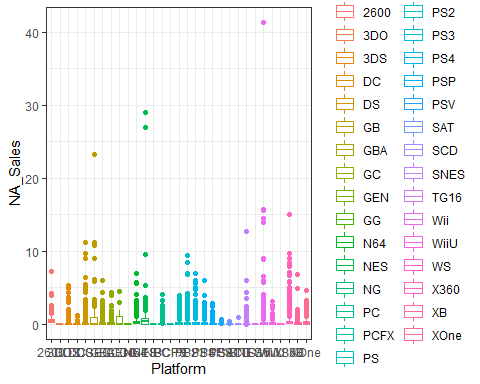
## Patform = Cualitativa

## NA\_Sales = Cuantitativa

## Graficamente = Boxplot

## Modelo = Anova

ggplot(data = VideoGames, aes(x =Platform, y=NA\_Sales, color=Platform))+ geom\_boxplot() + theme\_bw()



## Como vemos la grafica tiene una gran vialibidad de datos que tiene un comportamiento diferentes de tipo de plataforma, donde el tipo plataforma wii vemos donde su vialidad de datos mas elevado a los demas es decir es el que tiene mas ventas en norteamerica.

## Hipotesis

## No hay diferencia entre las medias del los grupos

## Al menor un par de medias son significativas distintas la una de la otra

anova<-aov(VideoGames$NA\_Sales~VideoGames$Platform)

summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## VideoGames$Platform 30 426 14.208 22.29 <2e-16 \*\*\*  
## Residuals 16688 10638 0.637   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Pr(>F)=2e-16 \*\*\*

## Respuesta: Debido a que el p-value es mayor que al nivel de significancia, no se rechza la hipoteis nula, por lo que no hay suficiente evidencia para cosiderar un par de medias son significativas distintas entre los tipos de plataforma.

## Pregunta 8:

## Realizar un analisis exploratorio para saber si existe alguna relacion entre en las ventas de videojuego en resto del mundo (Other\_Sales) y las ventas global (Global\_Sales).

## Other\_Sales = Cuantitativa

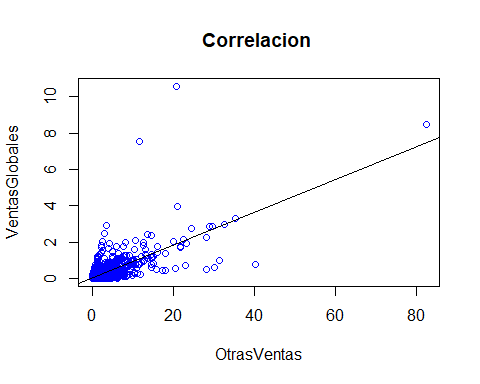
## Global\_Sales = Cuantitativa

## Graficamente = Grafico de dispersion (plot)

## Modelo = Correlacion

plot( Other\_Sales~Global\_Sales, data= VideoGames, col ="blue", main = "Correlacion ", xlab = "OtrasVentas", ylab = "VentasGlobales") + abline(lm(Other\_Sales~Global\_Sales, data=VideoGames, col="red"))

## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :  
## extra argument 'col' will be disregarded



## integer(0)

## Hipotesis

## A mayor ventas globales mayor es la OtrosVentras en el resto del mundo

## A mayor ventas globales menor es la OtrosVentras en el resto del mundo

cor(VideoGames$Other\_Sales,VideoGames$Global\_Sales)

## [1] 0.7492421

## Respuesta: Del resultado vemos que existe una correlacion de 0.7492421 entre las dos variables, esto significa no hay evidencia sufiente que nos indique que las variable esta correlacionda, por ende las ventas globales no van a mejorar mejor tal como las otras ventas en el resto del mundo.

## Pregunta 9:

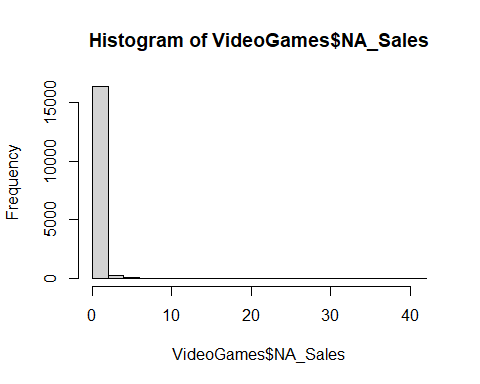
## Explorar la variable ventas en Norteámerica (NA\_Sales ) para ver la normlidad de los datos.

## NA\_Sales = cuantitativa

## Graficamente = histograma

## Modelo = shapiro

hist(VideoGames$NA\_Sales)



## Hipotesis

## Ho: Normalidad de los datos

## h1: No hay normalidad en los datos

ad.test(VideoGames$NA\_Sales)

##   
## Anderson-Darling normality test  
##   
## data: VideoGames$NA\_Sales  
## A = 3281.3, p-value < 2.2e-16

2.2e-16<0.05

## [1] TRUE

## Respuesta: Como el pvalue es mayo que el nivel de significancia 0.05, no se rechza la hipotesis Ho, hay normalida en los de las ventas en norteamerica.

## Pregunta 10:

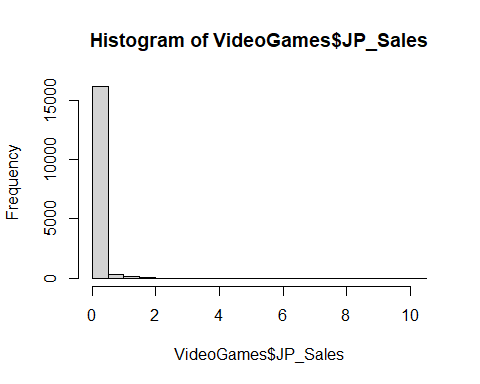
## Explorar la variable ventas en Japon (JP\_Sales ) para ver la normlidad de los datos.

## JP\_Sales = cuantitativa

## Graficamente = histograma

## Modelo = shapiro

hist(VideoGames$JP\_Sales)



## Hipotesis

## Ho: Normalidad de los datos

## h1: No hay normalidad en los datos

ad.test(VideoGames$JP\_Sales)

##   
## Anderson-Darling normality test  
##   
## data: VideoGames$JP\_Sales  
## A = 4109.5, p-value < 2.2e-16

2.2e-16<0.05

## [1] TRUE

## Respuesta: Como el p-value es mayo al nivel de significancia, no se rechaza la hipotesis Ho, por lo que hay normalidad de los datos de las ventas en Japon.