Trabajo EDA

Grupo 1

2022-09-25

# Análisis exploratorio de datos

## Explorando la base de datos

Para la realización del siguiente trabajo, hemos elegido emplear un dataset sobre episodios de la serie televisiva “Game of Thrones” que hemos extrído desde la siguiente dirección: <https://www.kaggle.com/datasets/rezaghari/game-of-thrones> . Esta base de datos refiere a información extraída principalmente del sitio IMBD a partir de webscrapping.

#Primero seteamos el directorio de trabajo  
library(here)  
  
#Ahora cargamos el archivo  
df = read.csv(here("Tareas","GOT\_episodes\_v4.csv"))  
  
  
#Seteamos las bibliotecas que emplearemos  
library(tidyverse)  
  
#Luego, damos un vistazo a la estructura de los datos  
library(dplyr)  
library(janitor)  
  
glimpse(df)

## Rows: 73  
## Columns: 18  
## $ Season <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, …  
## $ Episode <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1, 2, 3, 4, 5, 6, 7, 8,…  
## $ Title <chr> "Winter Is Coming", "The Kingsroad", "Lord Snow", "Cri…  
## $ Release\_date <chr> "17-Apr-11", "24-Apr-11", "1-May-11", "8-May-11", "15-…  
## $ Rating <dbl> 9.1, 8.8, 8.7, 8.8, 9.1, 9.2, 9.2, 9.0, 9.6, 9.5, 8.8,…  
## $ Votes <int> 38639, 29285, 27694, 26284, 27349, 27079, 27556, 25645…  
## $ Summary <chr> "Eddard Stark is torn between his family and an old fr…  
## $ Writer\_1 <chr> "David Benioff", "David Benioff", "David Benioff", "Da…  
## $ Writer\_2 <chr> "D.B. Weiss", "D.B. Weiss", "D.B. Weiss", "D.B. Weiss"…  
## $ Star\_1 <chr> "Sean Bean", "Sean Bean", "Sean Bean", "Sean Bean", "S…  
## $ Star\_2 <chr> "Mark Addy", "Mark Addy", "Mark Addy", "Mark Addy", "M…  
## $ Star\_3 <chr> "Nikolaj Coster-Waldau", "Nikolaj Coster-Waldau", "Nik…  
## $ Users\_reviews <int> 61, 27, 21, 22, 24, 23, 21, 20, 30, 34, 22, 19, 15, 19…  
## $ Critics\_reviews <int> 30, 28, 29, 26, 26, 26, 27, 26, 29, 29, 29, 25, 25, 24…  
## $ US\_Viewers <dbl> 2.22, 2.20, 2.44, 2.45, 2.58, 2.44, 2.40, 2.72, 2.66, …  
## $ Duration <int> 62, 56, 58, 56, 55, 53, 58, 59, 57, 53, 53, 54, 53, 51…  
## $ Director <chr> "Timothy Van Patten", "Timothy Van Patten", "Brian Kir…  
## $ Budget\_estimate <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …

Esta base de datos tiene 18 variables con 73 observaciones. Los tipos de variables que se encuentran son: interger, character y double. Así vemos que la base de datos cumple con los requisitos solicitados para su análisis.

Al verificar que existen variables de tipo character, procedemos a convertirlas en tipo factor:

df = df %>% mutate (Episode = as.factor(Episode),  
 Title = as.factor(Title),  
 Release\_date = as.factor(Release\_date),  
 Season= as.factor(Season),  
 Summary= as.factor(Summary),  
 Writer\_1= as.factor(Writer\_1),  
 Writer\_2= as.factor(Writer\_2),  
 Star\_1 = as.factor(Star\_1),  
 Star\_2 = as.factor(Star\_2),  
 Star\_3 = as.factor(Star\_3),  
 Director = as.factor(Director)  
 )  
  
glimpse(df)

## Rows: 73  
## Columns: 18  
## $ Season <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, …  
## $ Episode <fct> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1, 2, 3, 4, 5, 6, 7, 8,…  
## $ Title <fct> "Winter Is Coming", "The Kingsroad", "Lord Snow", "Cri…  
## $ Release\_date <fct> 17-Apr-11, 24-Apr-11, 1-May-11, 8-May-11, 15-May-11, 2…  
## $ Rating <dbl> 9.1, 8.8, 8.7, 8.8, 9.1, 9.2, 9.2, 9.0, 9.6, 9.5, 8.8,…  
## $ Votes <int> 38639, 29285, 27694, 26284, 27349, 27079, 27556, 25645…  
## $ Summary <fct> "Eddard Stark is torn between his family and an old fr…  
## $ Writer\_1 <fct> David Benioff, David Benioff, David Benioff, David Ben…  
## $ Writer\_2 <fct> D.B. Weiss, D.B. Weiss, D.B. Weiss, D.B. Weiss, D.B. W…  
## $ Star\_1 <fct> Sean Bean, Sean Bean, Sean Bean, Sean Bean, Sean Bean,…  
## $ Star\_2 <fct> Mark Addy, Mark Addy, Mark Addy, Mark Addy, Mark Addy,…  
## $ Star\_3 <fct> Nikolaj Coster-Waldau, Nikolaj Coster-Waldau, Nikolaj …  
## $ Users\_reviews <int> 61, 27, 21, 22, 24, 23, 21, 20, 30, 34, 22, 19, 15, 19…  
## $ Critics\_reviews <int> 30, 28, 29, 26, 26, 26, 27, 26, 29, 29, 29, 25, 25, 24…  
## $ US\_Viewers <dbl> 2.22, 2.20, 2.44, 2.45, 2.58, 2.44, 2.40, 2.72, 2.66, …  
## $ Duration <int> 62, 56, 58, 56, 55, 53, 58, 59, 57, 53, 53, 54, 53, 51…  
## $ Director <fct> Timothy Van Patten, Timothy Van Patten, Brian Kirk, Br…  
## $ Budget\_estimate <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …

Ahora que los tipos de variable son los necesitamos para trabajar, vamos a verificar la existencia o no de datos perdidos:

any(is.na(df))

## [1] FALSE

## Análisis univariado de variables

### Variables numéricas

Para el análisis de variables cuantitativas hemos elegido trabajar con Rating, Votes, US\_Viewers.

Comenzamos con el Rating:

df %>% summarise(Rating\_promedio=mean(Rating),Rating\_mediana=median(Rating))

## Rating\_promedio Rating\_mediana  
## 1 8.839726 8.9

df %>% summarise(Desvío\_estandar\_rating=sd(Rating),Coeficiente\_variación\_rating=sd(Rating)/mean(Rating),Rango\_intercuartílico\_rating=IQR(Rating))

## Desvío\_estandar\_rating Coeficiente\_variación\_rating  
## 1 0.931358 0.1053605  
## Rango\_intercuartílico\_rating  
## 1 0.7

df %>% summarise(Cuartiles\_rating=quantile(Rating))

## Cuartiles\_rating  
## 1 4.1  
## 2 8.7  
## 3 8.9  
## 4 9.4  
## 5 9.9

A partir de los datos de centro, se observa que se trata de una serie que, en general, tiene un buen rating considerando que se trata de una escala del uno al diez. Asimimos, se verifica que el promedio y la mediana son similares. No se calculó la moda debido a que se trata de una variable numérica continua, por lo que carece de sentido. Observando los datos de dispersión, se comprueba que no se trata de picos, sino que la mayoría de los episodios tiene ratings por encima de 8, por lo que podemos hablar de se trataría de una serie bien puntuada. Esto se complementa con los datos de posición, ya que más del 50 % de los episodios tiene puntuación entre 8,7 y 9,4.

Con respecto a los Votos, obtuvimos los siguientes datos:

df %>% summarise(Votos\_promedio=mean(Votes),Votos\_mediana=median(Votes))

## Votos\_promedio Votos\_mediana  
## 1 47789.51 29302

df %>% summarise(Desvío\_estandar\_votos=sd(Votes),Coeficiente\_variación\_votos=sd(Votes)/mean(Votes),Rango\_intercuartílico\_votos=IQR(Votes))

## Desvío\_estandar\_votos Coeficiente\_variación\_votos Rango\_intercuartílico\_votos  
## 1 44738.79 0.9361635 18473

df %>% summarise(Cuartiles\_votos=quantile(Votes))

## Cuartiles\_votos  
## 1 22223  
## 2 23734  
## 3 29302  
## 4 42207  
## 5 220581

Observamos que la media y la mediana no se corresponden, por lo que no se ve una distribución de los votos equitativa entre los episodios. Asimismo, la dispersión es relativamente alta en función de los valores trabajados, lo que se verifica con el coeficiente de variación. Con estos datos ya es posible visualizar que el máximo es un posible dato atípico y es esperable que haya más. Por último, realizamos la misma observación que en el punto anterior para la moda.

Por último, para US\_Viewers:

df %>% summarise(Viewers\_promedio=mean(US\_Viewers),Viewers\_mediana=median(US\_Viewers))

## Viewers\_promedio Viewers\_mediana  
## 1 6.447808 6.64

df %>% summarise(Desvío\_estandar\_Viewers=sd(US\_Viewers),Coeficiente\_variación\_viewers=sd(US\_Viewers)/mean(US\_Viewers),Rango\_intercuartílico\_viewers=IQR(US\_Viewers))

## Desvío\_estandar\_Viewers Coeficiente\_variación\_viewers  
## 1 2.827372 0.4385012  
## Rango\_intercuartílico\_viewers  
## 1 3.92

df %>% summarise(Cuartiles\_viewers=quantile(US\_Viewers))

## Cuartiles\_viewers  
## 1 2.20  
## 2 3.90  
## 3 6.64  
## 4 7.82  
## 5 13.61

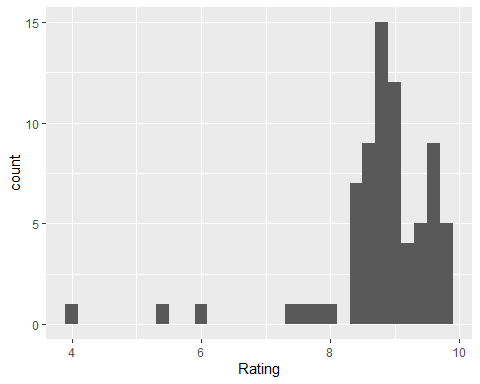
Se observa que el promedio y la mediana son similares, pero las medidas de dispersión refieren a que los datos no siguen el mismo patrón que el observado para el Rating. Por último, realizamos la misma observación que en el punto anterior para la moda.

Ahora pasaremos a realizar la visualización de las variables Rating y Viewers.

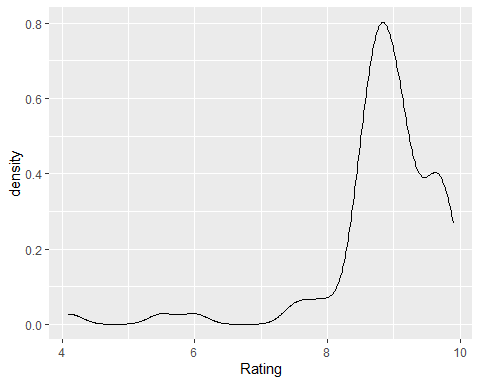
Rating:

df %>% ggplot()+geom\_histogram(aes(x=Rating))

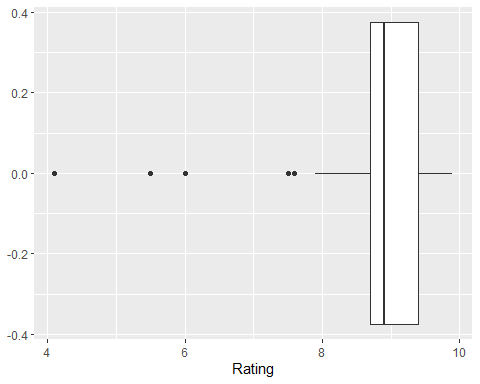
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



df %>% ggplot()+geom\_density(aes(x=Rating))



df %>% ggplot()+geom\_boxplot(aes(x=Rating))

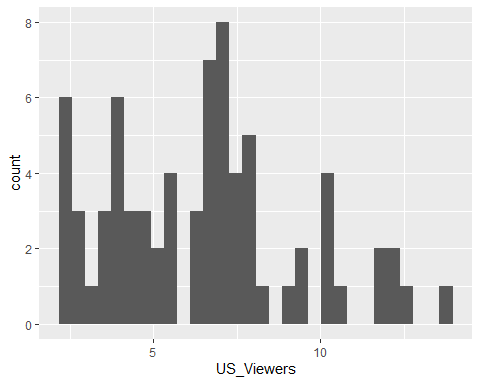


Para la variable Rating se observa un comportamiento bastante simétrico, lo que se refuerza con la similitud entre media y mediana. El boxplot revela la existencia de datos atípicos, vinculados a capítulos que tuvieron bajo rating por una baja aceptación del público. En la base de datos se observa que se trata de capítulos de la última temporada, la que fue altamente criticada de forma negativa por el público, vinculado a la alta expectativa del final de la serie.

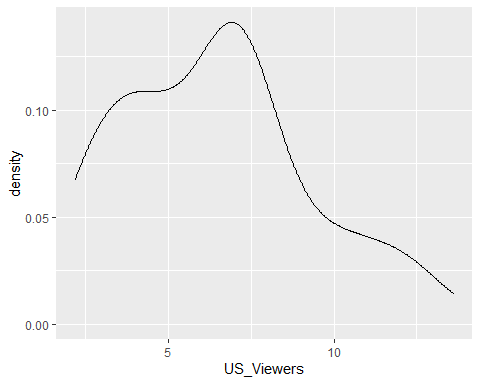
Viewers:

df %>% ggplot()+geom\_histogram(aes(x=US\_Viewers))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



df %>% ggplot()+geom\_density(aes(x=US\_Viewers))



df %>% ggplot()+geom\_boxplot(aes(x=US\_Viewers))



En relación a la variable Viewers se observa una distribución con una leve asimetría negativa. Por otra parte, el boxplot no revela datos atípicos a pesar de que existen varios valores relativamente grandes con respecto a la media.

### Variables categóricas

Para el análisis de las variables categóricas hemos decidido trabajar con los datos de Director y Stars.

Antes de comenzar, debemos mencionar que las variables Star\_1, Star\_2 y Star\_3 indican los actores que protagonizaron dicho episodio ordenados por cantidad de apariciones. Es por este motivo que decimos generar una nueva variable que contenga dichos datos ya que para un mismo episodio no se repetirá el dato de aparición de un actor y esto nos permitirá analizar qué actores aparecieron más veces, dato que no es posible obtener analizando por separado las variables Star\_1 o Star\_2 o Star\_3.

Procedemos a crear la variable Star en un nuevo dataframe y analizarla:

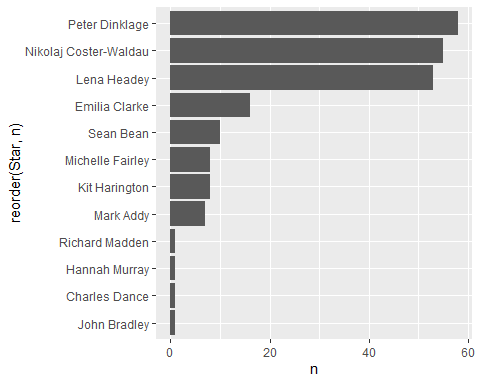
Star = c(df$Star\_1,df$Star\_2,df$Star\_3)  
  
Star = as.data.frame(Star)  
  
Star %>% count(Star) %>% mutate(prop=n/(sum(n)/3))

## Star n prop  
## 1 Emilia Clarke 16 0.21917808  
## 2 Kit Harington 8 0.10958904  
## 3 Nikolaj Coster-Waldau 55 0.75342466  
## 4 Peter Dinklage 58 0.79452055  
## 5 Sean Bean 10 0.13698630  
## 6 John Bradley 1 0.01369863  
## 7 Lena Headey 53 0.72602740  
## 8 Mark Addy 7 0.09589041  
## 9 Michelle Fairley 8 0.10958904  
## 10 Charles Dance 1 0.01369863  
## 11 Hannah Murray 1 0.01369863  
## 12 Richard Madden 1 0.01369863

#Moda  
names(which.max(table(Star)))

## [1] "Peter Dinklage"

tabla\_star = Star %>% count(Star) %>% mutate(prop=n/73)  
tabla\_star %>% ggplot() + geom\_col(aes(x=n,y=reorder(Star,n)))



A partir del análisis de la variable Star, es posible visualizar que el actor que participó de la mayor parte de los capítulos fue Peter Dinklage con una participación de un 79,5% de los mismos, siendo seguido por Nikolaj Coster-Waldau y Lena Headey. Por último, lo que podemos afirmar con seguridad es la alta participación de estos personajes, no así de los restantes por la información contenida en la base de datos y la forma en que fue construída la variable.

Para la variable director:

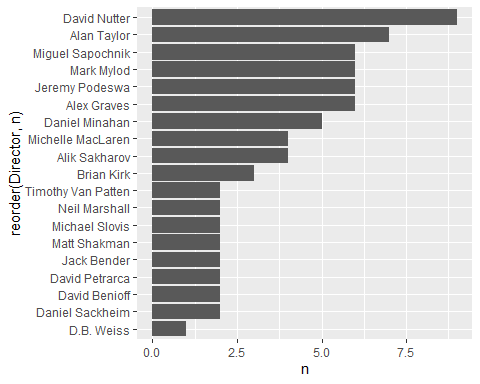
df %>% count(Director) %>% mutate(prop=n/sum(n)) %>% adorn\_totals()

## Director n prop  
## Alan Taylor 7 0.09589041  
## Alex Graves 6 0.08219178  
## Alik Sakharov 4 0.05479452  
## Brian Kirk 3 0.04109589  
## D.B. Weiss 1 0.01369863  
## Daniel Minahan 5 0.06849315  
## Daniel Sackheim 2 0.02739726  
## David Benioff 2 0.02739726  
## David Nutter 9 0.12328767  
## David Petrarca 2 0.02739726  
## Jack Bender 2 0.02739726  
## Jeremy Podeswa 6 0.08219178  
## Mark Mylod 6 0.08219178  
## Matt Shakman 2 0.02739726  
## Michael Slovis 2 0.02739726  
## Michelle MacLaren 4 0.05479452  
## Miguel Sapochnik 6 0.08219178  
## Neil Marshall 2 0.02739726  
## Timothy Van Patten 2 0.02739726  
## Total 73 1.00000000

#Moda  
names(which.max(table(df$Director)))

## [1] "David Nutter"

tabla\_director = df %>% count(Director) %>% mutate(prop=n/sum(n))  
tabla\_director %>% ggplot() + geom\_col(aes(x=n,y=reorder(Director,n)))



En el análisis de la variable Director observamos que el que participó de la producción de la mayor cantidad de capítulos fue David Nutter. También es posible observar que no existió un director que participará de la producción de muchos capítulos, sino que muchos directores realizaron pocos capítulos.

##Tabla de Contingencias

Para generar la tabla de contigencias, elegimos como variables categóricas a la variable Season y a los actores que participan en cada episodio (variable “Star”).

X=df$Director  
Y=df$Season  
  
table(X,Y) %>% addmargins()

## Y  
## X 1 2 3 4 5 6 7 8 Sum  
## Alan Taylor 2 4 0 0 0 0 1 0 7  
## Alex Graves 0 0 2 4 0 0 0 0 6  
## Alik Sakharov 0 1 1 2 0 0 0 0 4  
## Brian Kirk 3 0 0 0 0 0 0 0 3  
## D.B. Weiss 0 0 0 1 0 0 0 0 1  
## Daniel Minahan 3 0 2 0 0 0 0 0 5  
## Daniel Sackheim 0 0 0 0 0 2 0 0 2  
## David Benioff 0 0 1 0 0 0 0 1 2  
## David Nutter 0 2 2 0 2 0 0 3 9  
## David Petrarca 0 2 0 0 0 0 0 0 2  
## Jack Bender 0 0 0 0 0 2 0 0 2  
## Jeremy Podeswa 0 0 0 0 2 2 2 0 6  
## Mark Mylod 0 0 0 0 2 2 2 0 6  
## Matt Shakman 0 0 0 0 0 0 2 0 2  
## Michael Slovis 0 0 0 0 2 0 0 0 2  
## Michelle MacLaren 0 0 2 2 0 0 0 0 4  
## Miguel Sapochnik 0 0 0 0 2 2 0 2 6  
## Neil Marshall 0 1 0 1 0 0 0 0 2  
## Timothy Van Patten 2 0 0 0 0 0 0 0 2  
## Sum 10 10 10 10 10 10 7 6 73

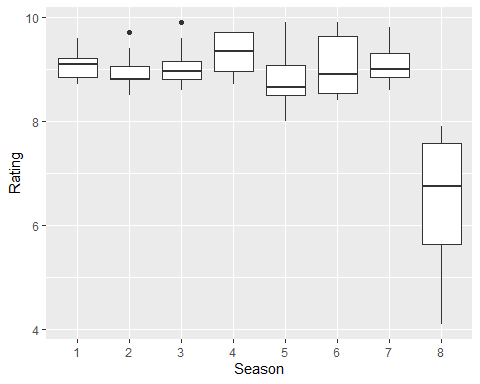
Queremos ver luego cómo se comporta la variable numérica de rating en base a la variable categórica de Temporada (Season) que utilizamos en el punto anterior.

df %>%   
 group\_by(Season) %>%  
 summarise(media = mean(Rating),  
 mediana = median(Rating),  
 desv\_est = sd(Rating),  
 Coef\_Var = sd(Rating)/mean(Rating)  
 )

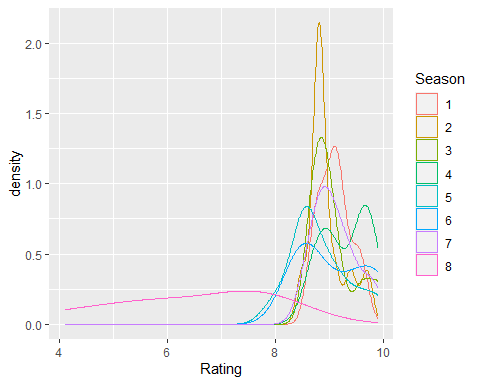
## # A tibble: 8 × 5  
## Season media mediana desv\_est Coef\_Var  
## <fct> <dbl> <dbl> <dbl> <dbl>  
## 1 1 9.1 9.1 0.294 0.0324  
## 2 2 8.96 8.8 0.350 0.0391  
## 3 3 9.05 8.95 0.412 0.0455  
## 4 4 9.3 9.35 0.419 0.0451  
## 5 5 8.83 8.65 0.556 0.0630  
## 6 6 9.06 8.9 0.620 0.0685  
## 7 7 9.1 9 0.404 0.0444  
## 8 8 6.43 6.75 1.49 0.232

Luego, procedemos a realizar un boxplot y un density plot de las variables elegidas en el punto anterior:

ggplot(data=df) + geom\_boxplot(aes(x=Season,y=Rating))



ggplot(data=df) + geom\_density(aes(x=Rating,color=Season))



#df %>% ggplot()+geom\_histogram(aes(x=Rating,color=Season))

En este punto calcularemos la correlación lineal entre las variables numéricas del dataset:

cor(df %>% select(Rating,Votes,Users\_reviews,Critics\_reviews,US\_Viewers,Duration))

## Rating Votes Users\_reviews Critics\_reviews US\_Viewers  
## Rating 1.0000000 -0.5287481 -0.8137918 -0.5373986 -0.4653055  
## Votes -0.5287481 1.0000000 0.7833720 0.7072795 0.6573201  
## Users\_reviews -0.8137918 0.7833720 1.0000000 0.5588094 0.5345244  
## Critics\_reviews -0.5373986 0.7072795 0.5588094 1.0000000 0.8772556  
## US\_Viewers -0.4653055 0.6573201 0.5345244 0.8772556 1.0000000  
## Duration -0.5196508 0.6613801 0.6789899 0.5275561 0.5446156  
## Duration  
## Rating -0.5196508  
## Votes 0.6613801  
## Users\_reviews 0.6789899  
## Critics\_reviews 0.5275561  
## US\_Viewers 0.5446156  
## Duration 1.0000000

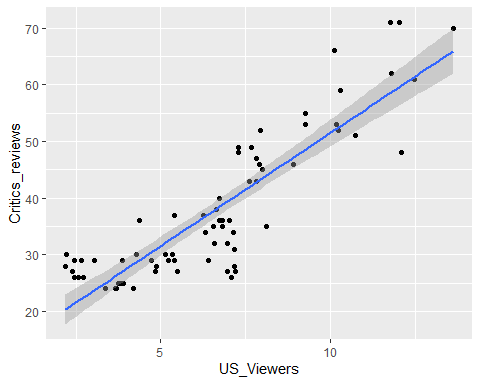
Vemos que la variable Rating tiene correlación negativa con el resto de las variables, mientras que el resto de las variables tienen correlaciones positivas entre sí. En general, la mayoría de las correlación son moderadas, destacando Rating vs Users\_reviews y US\_Viewers vs Critics\_reviews como fuertes.

Vamos a analizar esta última correlación: US\_Viewers vs Critics\_reviews por tratarse de la relación con la mayor correlación lineal.

lm(Critics\_reviews~US\_Viewers,df)

##   
## Call:  
## lm(formula = Critics\_reviews ~ US\_Viewers, data = df)  
##   
## Coefficients:  
## (Intercept) US\_Viewers   
## 11.612 3.991

df %>% ggplot(aes(x=US\_Viewers,y=Critics\_reviews)) + geom\_point() + geom\_smooth(method = "lm")



Podemos hablar de que existe una posible relación lineal entre la cantidad de visualizaciones de la serie respecto a la cantidad de reviews realizadas por la crítica especializada sobre la serie. Tiene sentido debido a que a medida que la gente se interesaba más en mirar la serie, la crítica especializada se interesaba más en escribir reseñas sobre la misma, especialmente para hacer llegar su trabajo a dicho público. A partir del modelo lineal descrito, se observa que por cada millón de televidentes se incrementaba en casi cuatro la cantidad de reseñas y que el piso de reseñas por capítulo es de once.