Trabajo EDA

Grupo 1: Martín Botta - Nicolás Gómez - Germán Miranda

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.

```
#Seteamos las bibliotecas que emplearemos
library(tidyverse)
library(dplyr)
library(janitor)
library(kableExtra)
library(here)

#Utilizamos la libreria here para localizar el directorio de trabajo y leer el archivo

df = read.csv(here("Tareas","GOT_episodes_v4.csv"))

#Luego, damos un vistazo a la estructura de los datos
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, ~
```

Descripción de Variables

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.

Variable	**Tipo**	**Descripción**
Season	Categórica ordinal	Número de temporada
Episode	Categórica ordinal	Número de episodio
Title	Categórica nominal	Título del episodio
Release_date	Categórica nominal	Fecha de lanzamiento del episodio
Rating	numérica continua	Rating obtenido en el sitio IMDB
Votes	numérica continua	Cantidad de votos
Summary	categórica nominal	Resumen del capítulo
Writer_1	categórica nominal	Primer escritor relacionado al capítulo
Writer_ 2	categórica nominal	Segundo escritor relacionado al capítulo
Star_1	categórica nominal	Primer actor relacionado al capítulo
Star_2	categórica nominal	Segundo actor relacionado al capítulo
Star_3	categórica nominal	Tercer actor relacionado al capítulo
User_reviews	numérica continua	Cantidad de usuarios que hicieron reseñas del capítulo
Critics_reviews	Numérica continua	Cantidad de críticos que hicieron reseñas del capítulo
US_Viewers	Numérica continua	Cantidad de visualizaciones en Estados Unidos
Duration	Numérica continua	Duración en minutos del capítulo
Director	Categórica nominal	Director del capítulo
Budget_estimate	Numérica continua	Estimado del presupuesto del capítulo

Table 1: Descripción de variables

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

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)
```

```
##   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
```

```
##   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)/mea
```

```
##   Desvío_estandar_Viewers Coeficiente_variación_viewers
## 1                2.827372                     0.4385012
##   Rango_intercuartílico_viewers
## 1                          3.92
```

```
df %>% summarise(Cuartiles_viewers=quantile(US_Viewers))
```

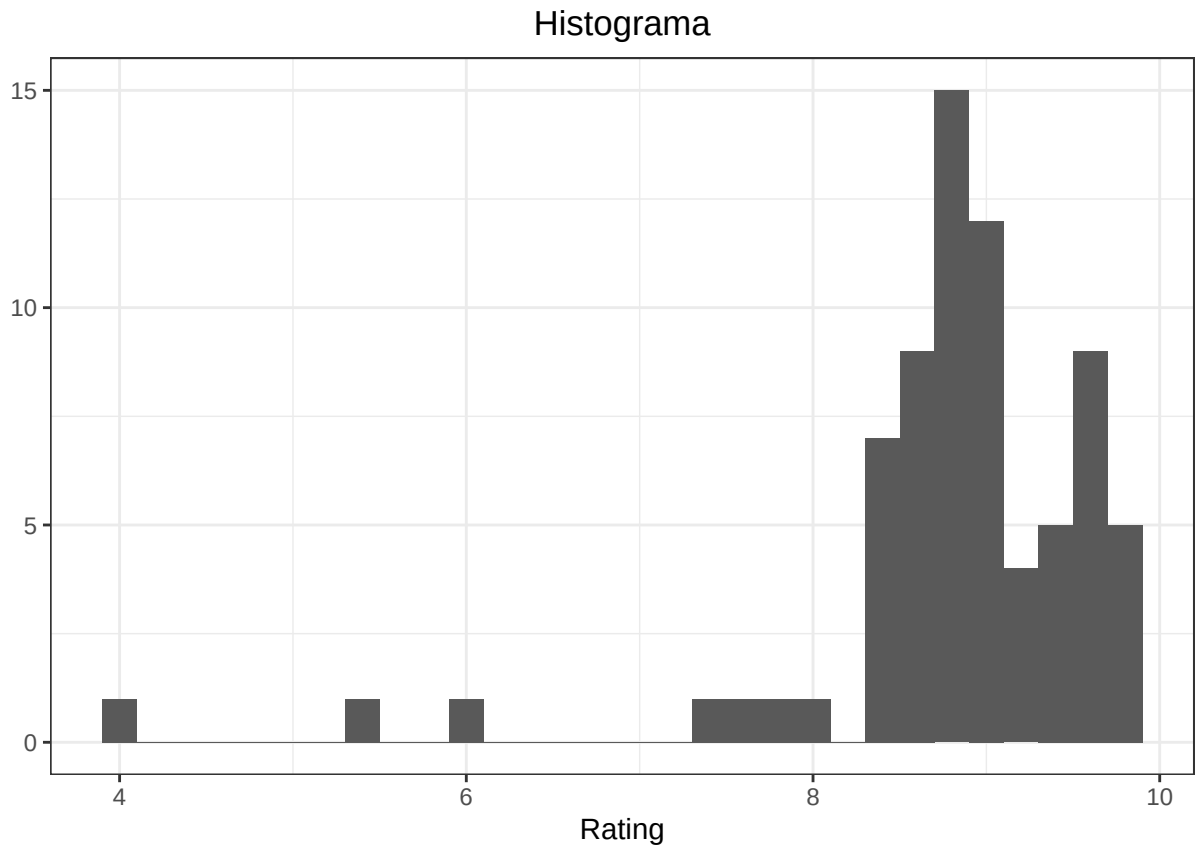
```
##   Cuartiles_viewers
## 1              2.20
## 2              3.90
## 3              6.64
## 4              7.82
## 5             13.61
```

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.

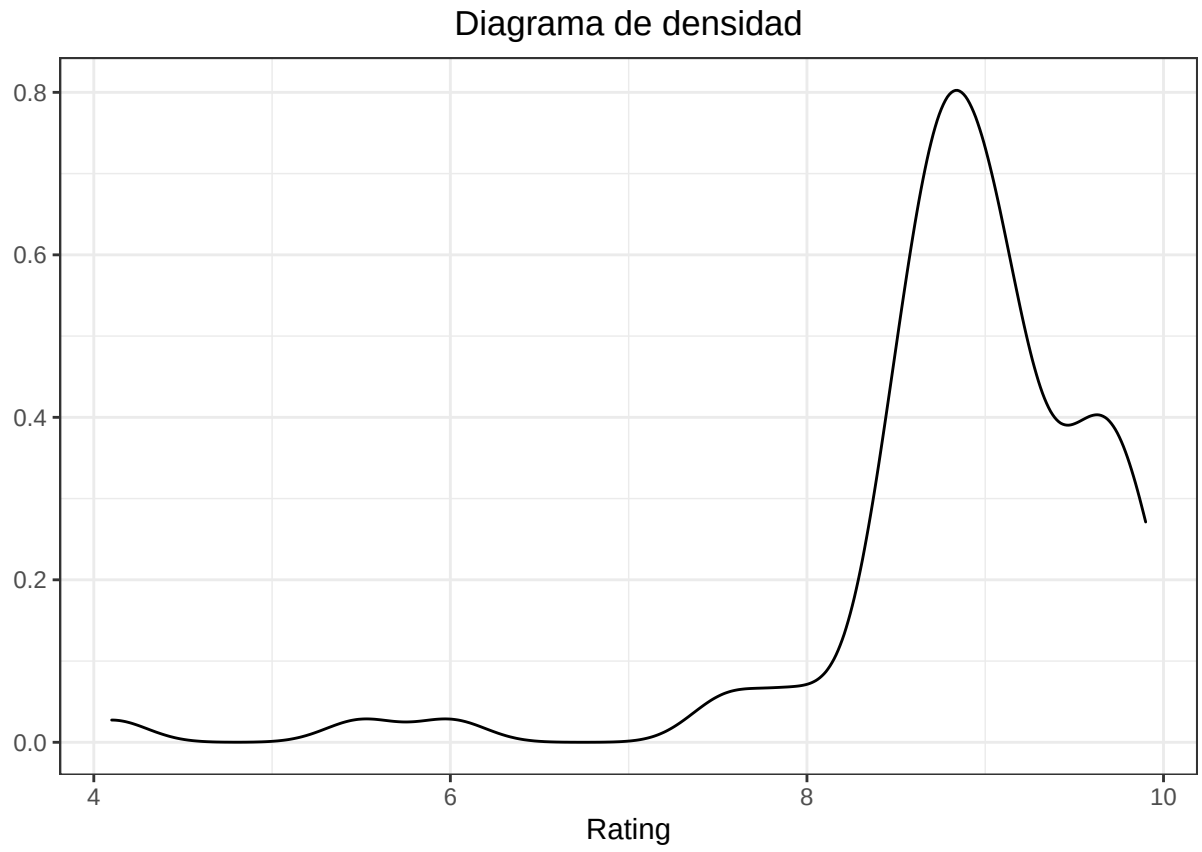
Ahora pasaremos a realizar la visualización de las variables Rating y Viewers.

Rating:

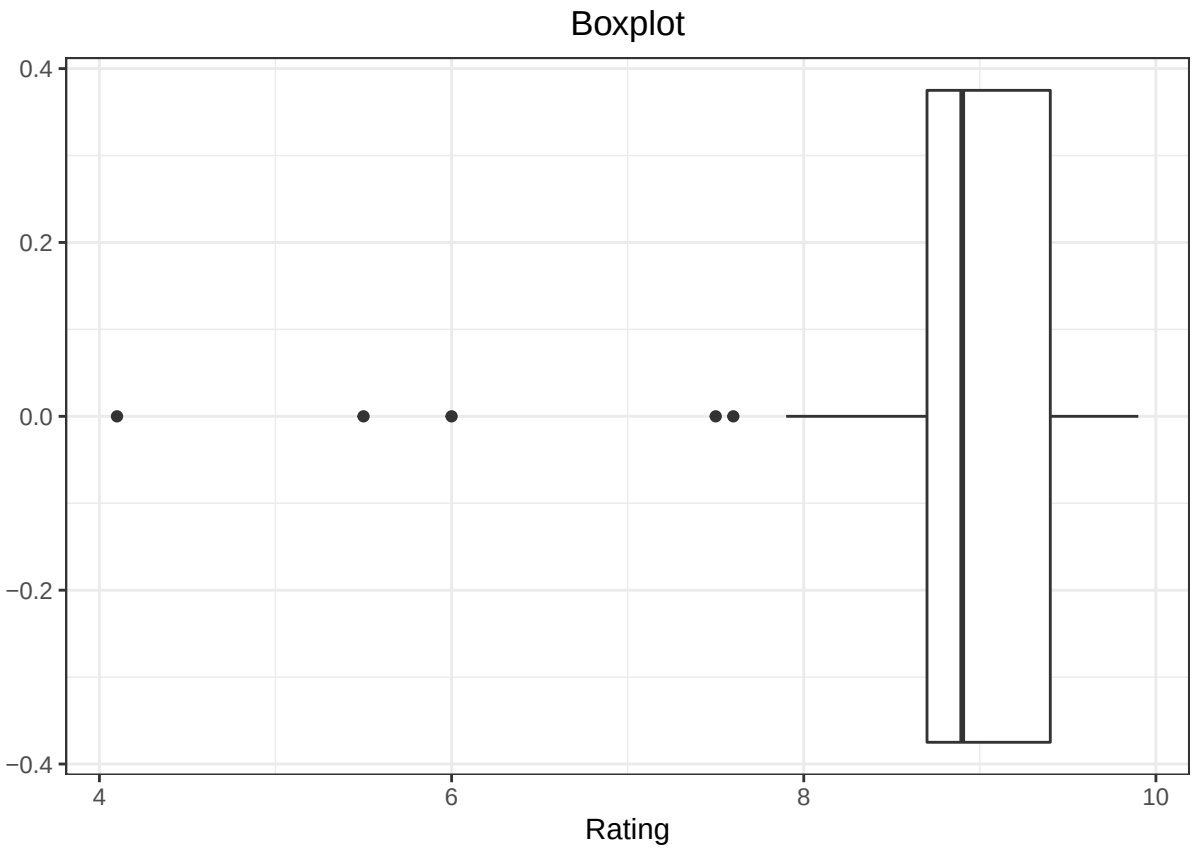
```
df %>% ggplot()+geom_histogram(aes(x=Rating))+labs(title="Histograma",)+ylab("")+theme_bw()+theme(plot.
```




```
df %>% ggplot()+geom_density(aes(x=Rating))+labs(title="Diagrama de densidad",)+ylab("")+theme_bw()+ther
```



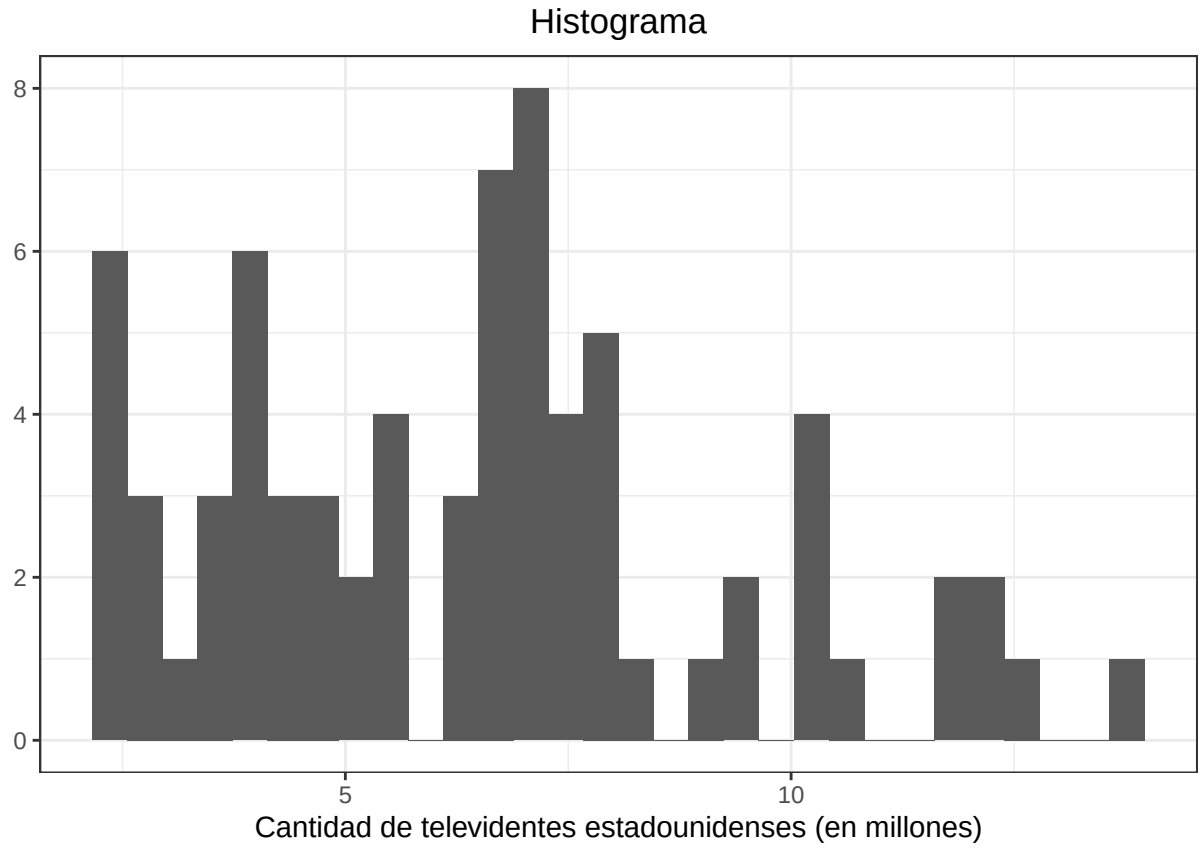

```
df %>% ggplot()+geom_boxplot(aes(x=Rating))+labs(title="Boxplot",)+ylab("")+theme_bw()+theme(plot.title=
```



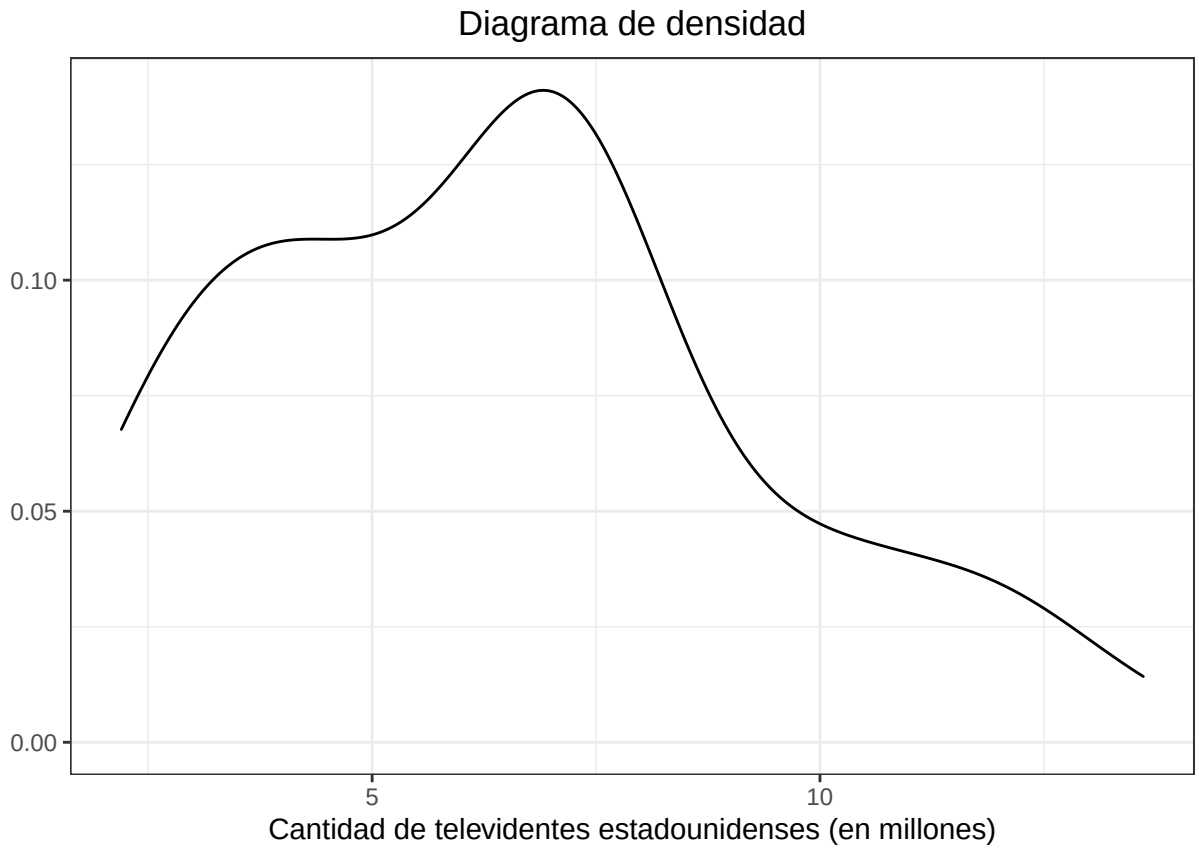
Para la variable Rating se observa un comportamiento bastante simétrico, lo que se refuerza con la similitud entre media y mediana. El boxplot revela la existencia de datos atípicos, vinculados a capítulos que tuvieron bajo rating por una baja aceptación del público. En la base de datos se observa que 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:

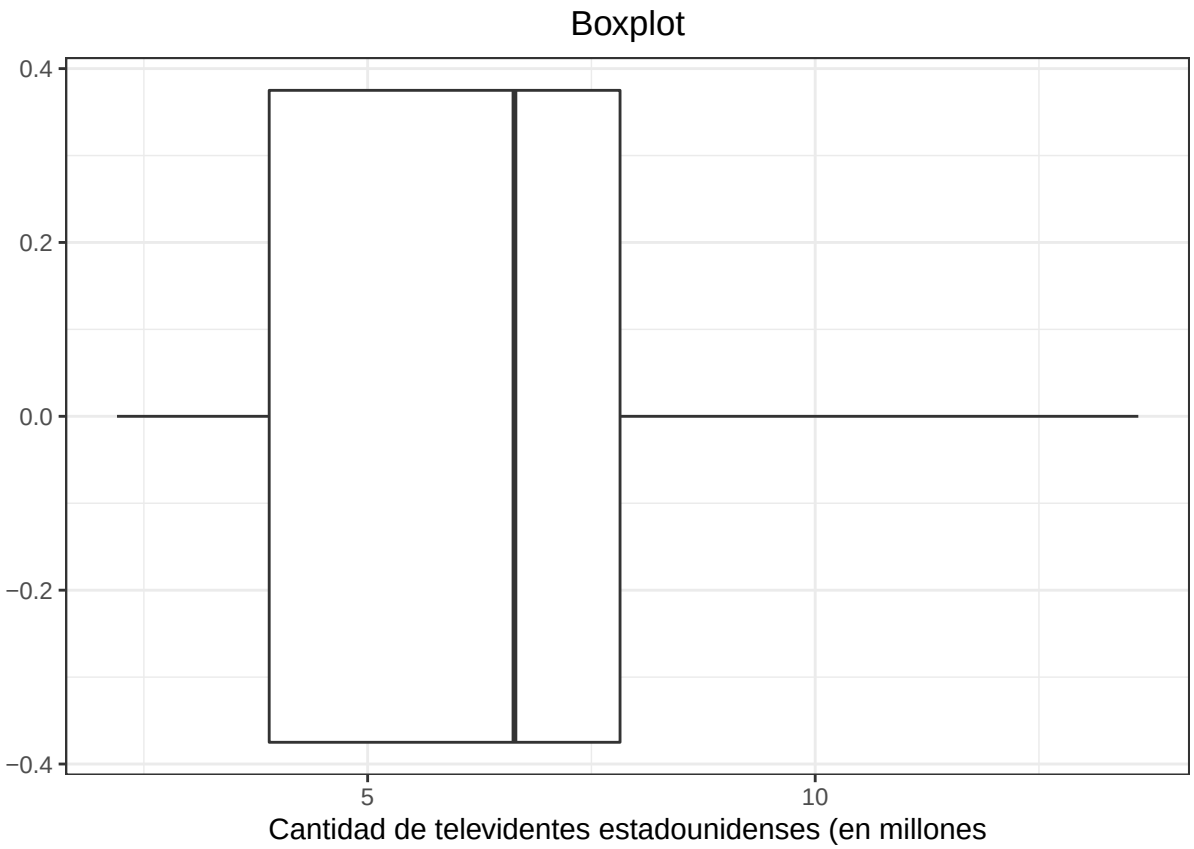
```
df %>% ggplot()+geom_histogram(aes(x=US_Viewers))+labs(title="Histograma",)+xlab("Cantidad de televident
```




```
df %>% ggplot()+geom_density(aes(x=US_Viewers))+labs(title="Diagrama de densidad",)+xlab("Cantidad de t
```




```
df %>% ggplot()+geom_boxplot(aes(x=US_Viewers))+labs(title="Boxplot",)+ylab("")+xlab("Cantidad de televi
```



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:

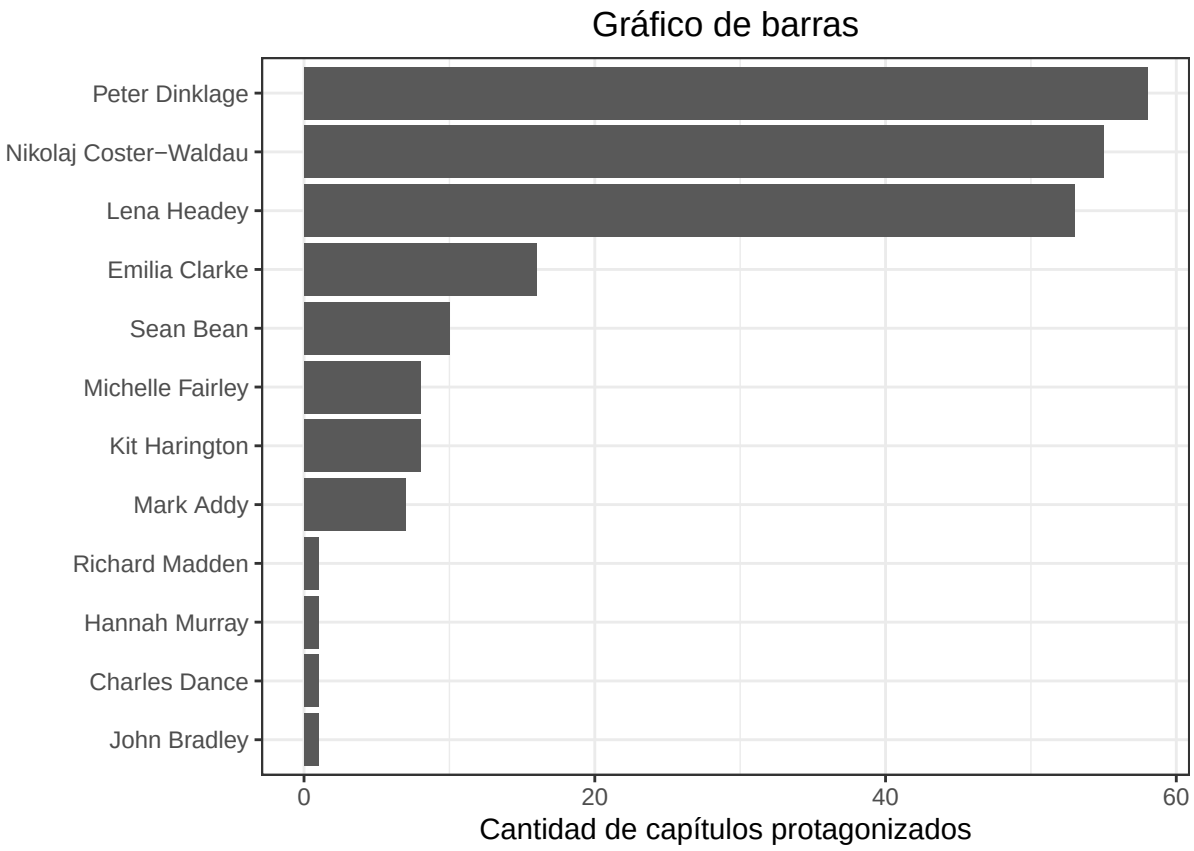
```
Star = c(df$Star_1,df$Star_2,df$Star_3)

Star = as.data.frame(Star)

tabla0 = Star %>% count(Star) %>% mutate(prop=n/(sum(n)/3))

kable(tabla0)
```

Star	n	prop
Emilia Clarke	16	0.2191781
Kit Harington	8	0.1095890
Nikolaj Coster-Waldau	55	0.7534247
Peter Dinklage	58	0.7945205
Sean Bean	10	0.1369863
John Bradley	1	0.0136986
Lena Headey	53	0.7260274
Mark Addy	7	0.0958904
Michelle Fairley	8	0.1095890
Charles Dance	1	0.0136986
Hannah Murray	1	0.0136986
Richard Madden	1	0.0136986

```
#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))) +labs(title="Gráfico de barras",)+ylab("
```



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:

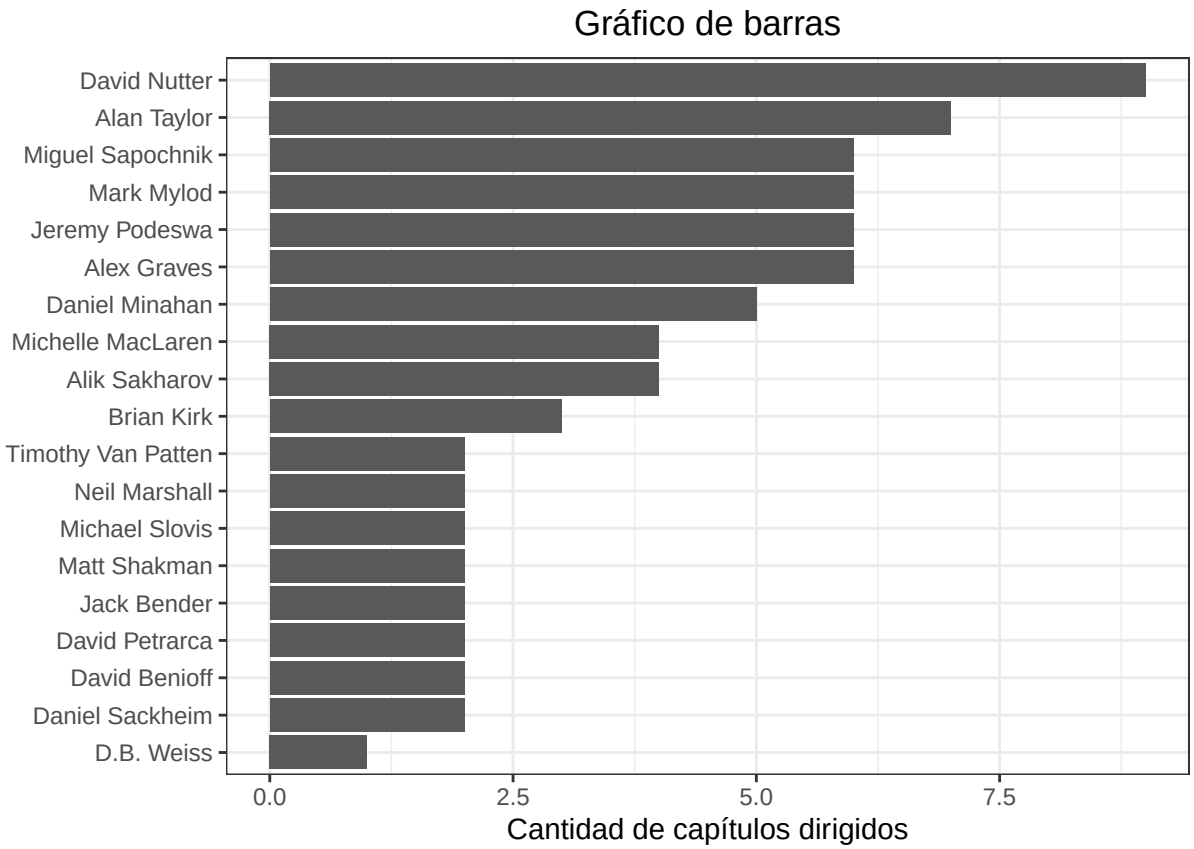
```
tabla1 = df %>% count(Director) %>% mutate(prop=n/sum(n)) %>% adorn_totals()

kable(tabla1)
```

Director	n	prop
Alan Taylor	7	0.0958904
Alex Graves	6	0.0821918
Alik Sakharov	4	0.0547945
Brian Kirk	3	0.0410959
D.B. Weiss	1	0.0136986
Daniel Minahan	5	0.0684932
Daniel Sackheim	2	0.0273973
David Benioff	2	0.0273973
David Nutter	9	0.1232877
David Petrarca	2	0.0273973
Jack Bender	2	0.0273973
Jeremy Podeswa	6	0.0821918
Mark Mylod	6	0.0821918
Matt Shakman	2	0.0273973
Michael Slovis	2	0.0273973
Michelle MacLaren	4	0.0547945
Miguel Sapochnik	6	0.0821918
Neil Marshall	2	0.0273973
Timothy Van Patten	2	0.0273973
Total	73	1.0000000

```
#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))) +labs(title="Gráfico de barras",
```



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.

Análisis multivariado

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

tabla2 = table(X,Y) %>% addmargins()

kable(tabla2)
```

	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.

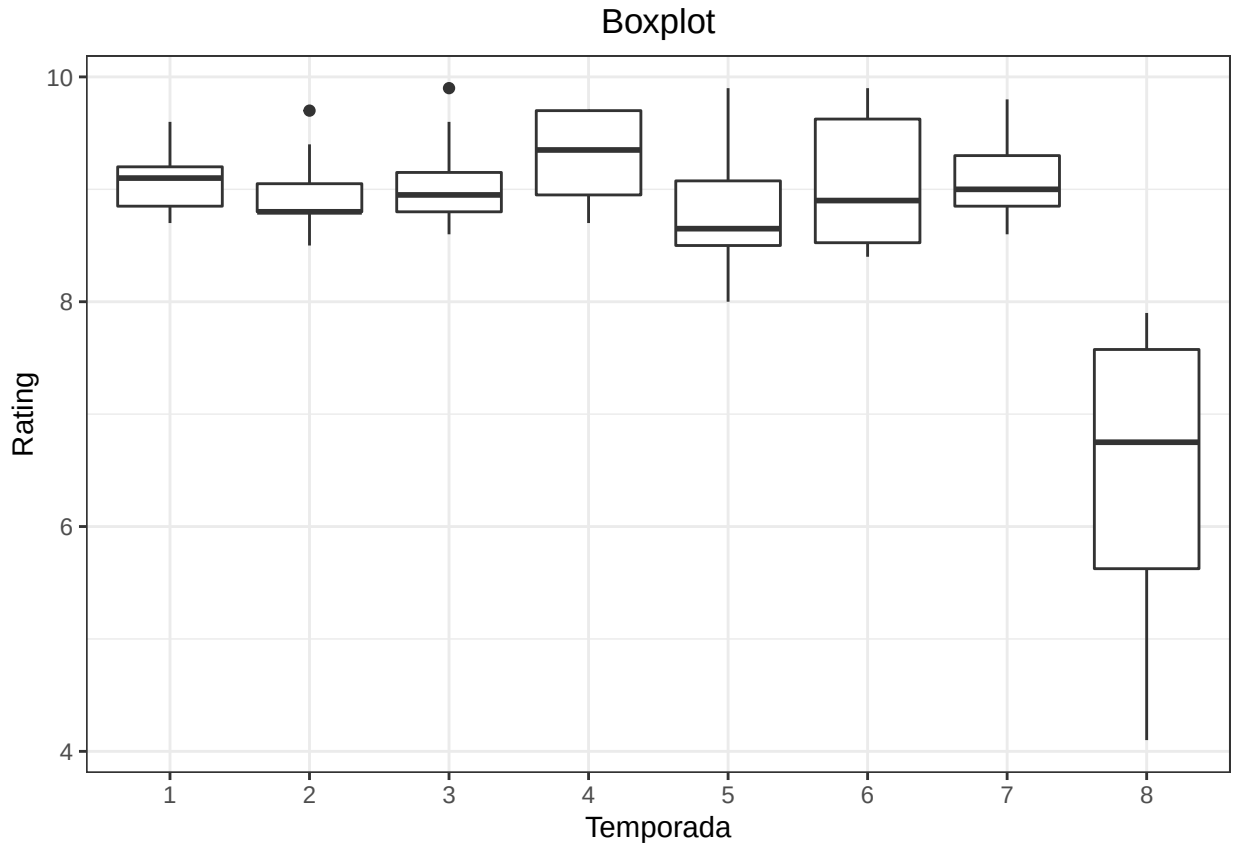
```
df %>%
  group_by(Season) %>%
    summarise(media = mean(Rating),
              mediana = median(Rating),
              desv_est = sd(Rating),
              Coef_Var = sd(Rating)/mean(Rating)
              )
```

```
## # A tibble: 8 x 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
```

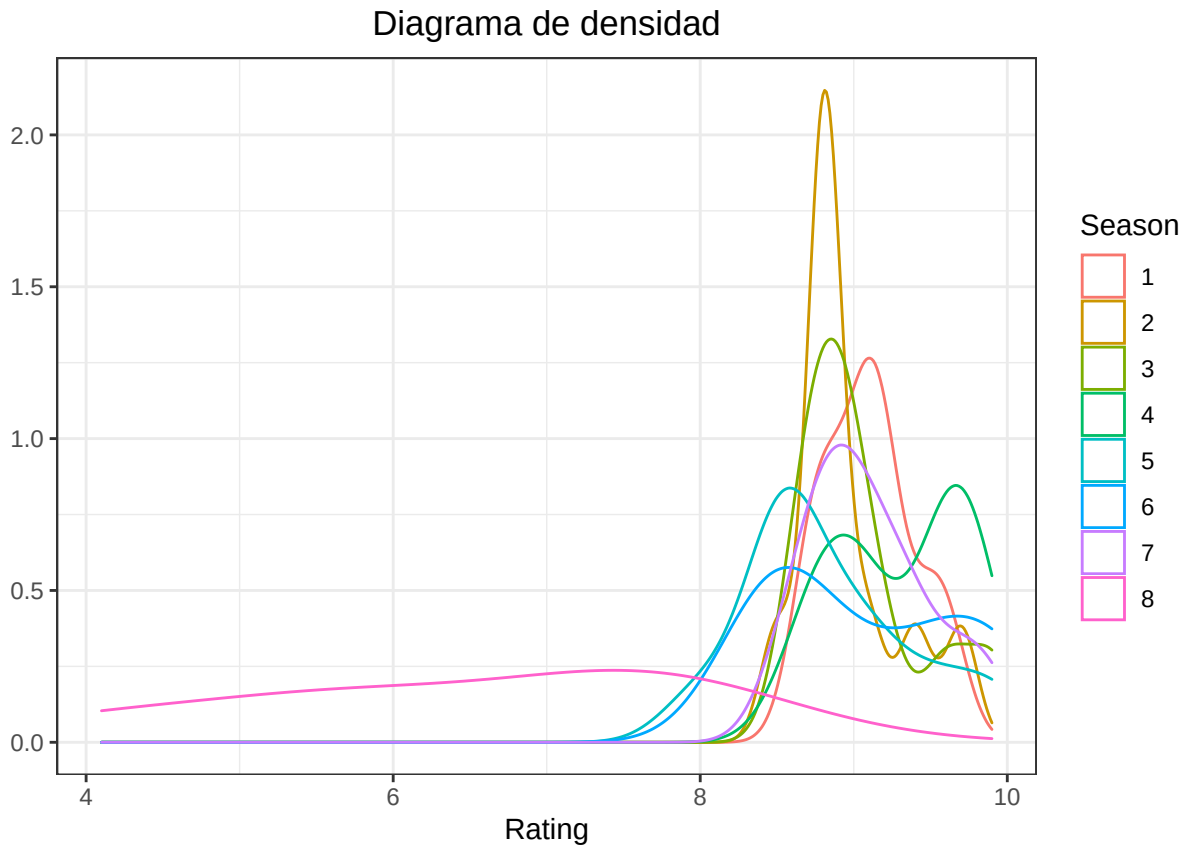
```
## 4 4       9.3     9.35    0.419   0.0451
## 5 5       8.83    8.65    0.556   0.0630
## 6 6       9.06    8.9     0.620   0.0685
## 7 7       9.1     9       0.404   0.0444
## 8 8       6.43    6.75    1.49    0.232
```

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

```
df %>% ggplot() + geom_boxplot(aes(x=Season,y=Rating))+labs(title="Boxplot",)+xlab("Temporada")+theme_bu
```



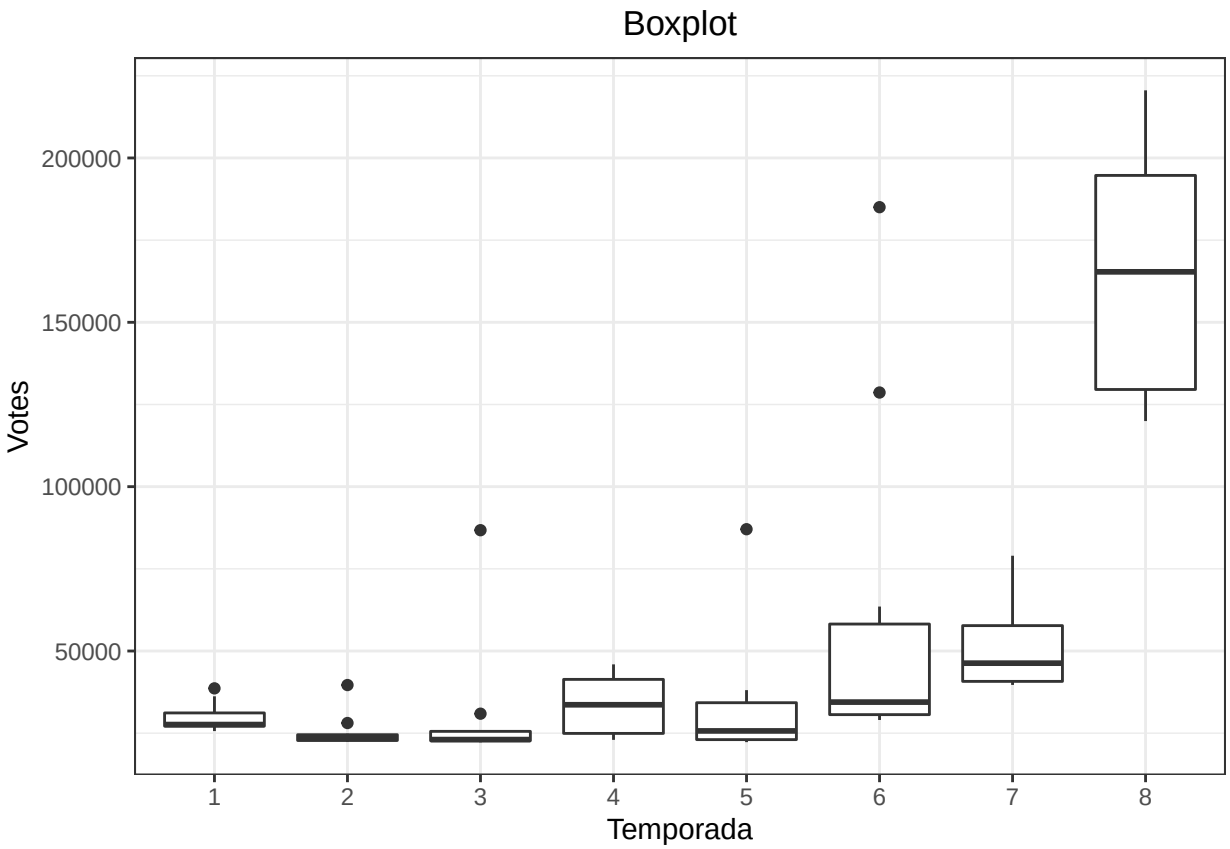

```
df %>% ggplot() + geom_density(aes(x=Rating,color=Season))+labs(title="Diagrama de densidad",)+ylab("")+
```



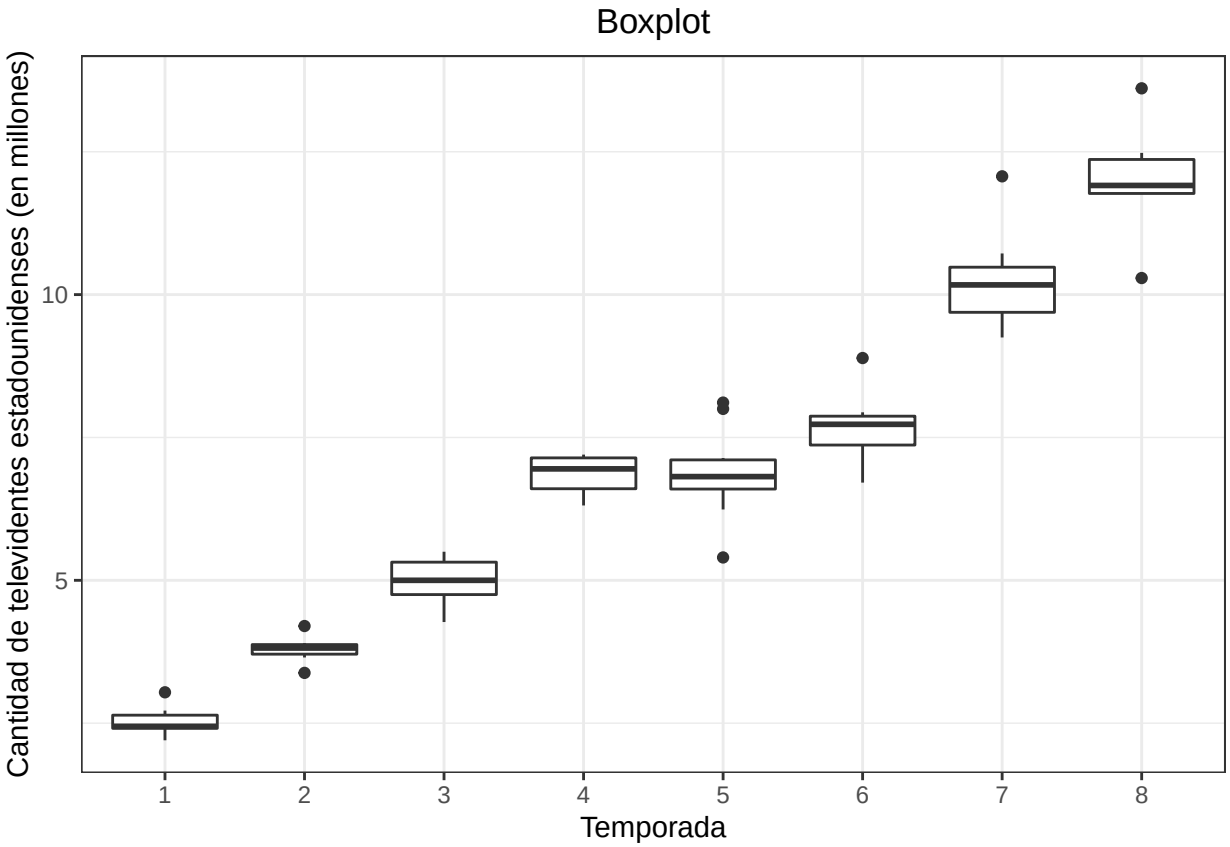

```
#df %>% ggplot()+geom_histogram(aes(x=Rating,color=Season))
```

Por otra parte, nos interesa conocer la distribución de los votos y de la cantidad de televidentes por temporada, por lo que procederemos a realizar dos boxplot:

```
df %>% ggplot() + geom_boxplot(aes(x=Season,y=Votes))+labs(title="Boxplot",)+xlab("Temporada")+theme_bw
```




```
df %>% ggplot() + geom_boxplot(aes(x=Season,y=US_Viewers))+labs(title="Boxplot",)+xlab("Temporada")+ylab
```



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
```

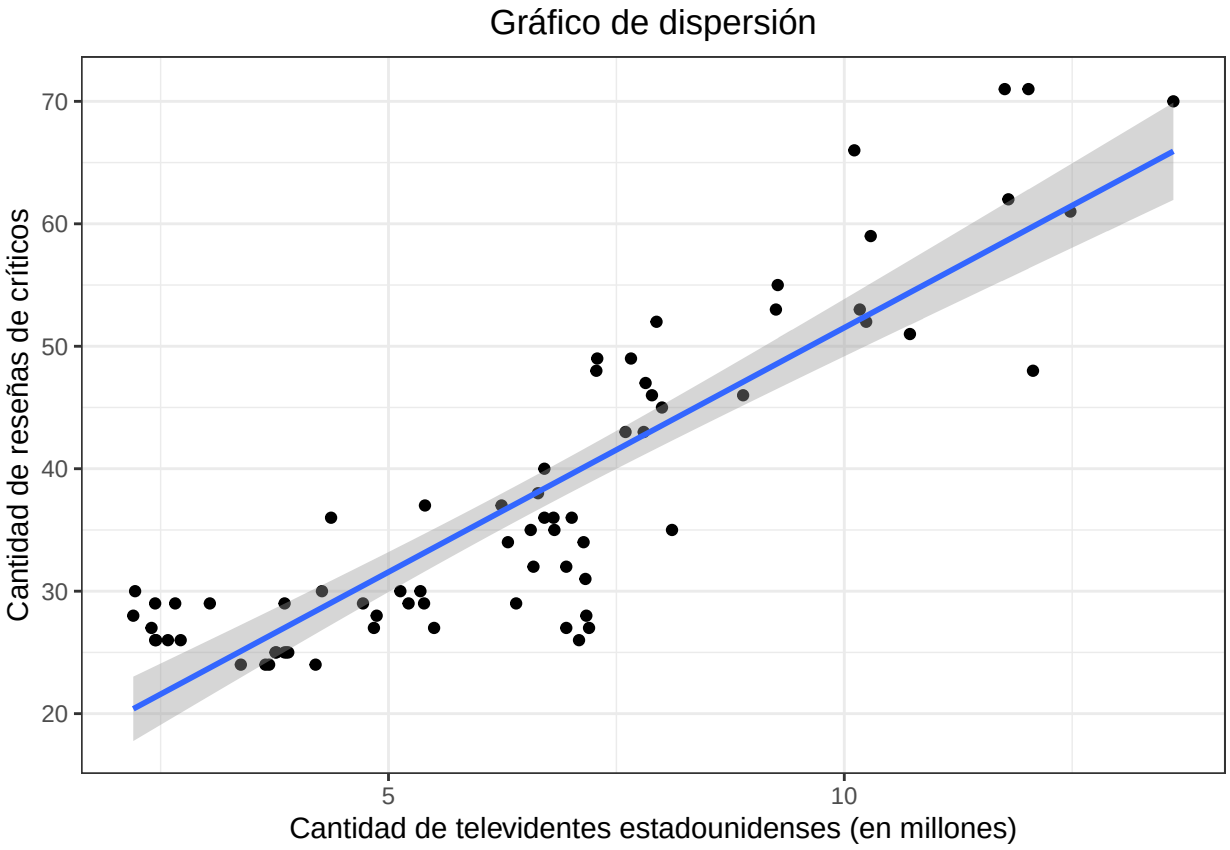
Vemos que la variable Rating tiene correlación negativa con el resto de las variables, mientras que el resto de las variables tienen correlaciones positivas entre sí. En general, la mayoría de las correlación son moderadas, destacando Rating vs Users_reviews y US_Viewers vs Critics_reviews como fuertes.

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")+labs(tit
```



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 poco menos de doce.

Conclusiones

Observamos que la serie, en general, tiene buena puntuación, sobre todo considerando que los capítulos tienen ratings por encima de los 8 puntos, excepto por capítulos de la última temporada. Así, la temporada cuatro es la mejor valorada, con capítulos, en general, con rating por encima de nueve, mientras que la temporada dos es la más homogénea en rating, debido a que la mayoría de los capítulos tienen rating alrededor de 8,8.

A pesar de esto, la temporada 8 presenta capítulos con valores de rating muy por debajo a la media, con el valor mínimo de 4,1 para el final de la serie. Esto podría estar vinculado a expectativas no cumplidas por la serie de sus espectadores.

Asimismo, si bien la cantidad de votos para las temporadas uno a siete se mantiene estable, para la temporada ocho se dispara. Entendemos que puede tratarse a estas expectativas no cumplidas por la serie, más que a un uso extensivo de la plataforma de donde se obtienen los datos. Esto último es un elemento a considerar debido a las fechas en que fueron producidos los capítulos, sin embargo nos inclinamos a pensar que esta última temporada no supo cumplir las demandas de los espectadores.

También se observa un crecimiento de la cantidad de televidentes a medida que fueron pasando las diferentes temporadas, que entendemos se debe a la popularidad que fue ganando la serie en la medida que fue siendo publicitada y viralizada.

Un elemento que nos llamó la atención es que existe una gran cantidad de directores, y que una misma temporada es dirigida por múltiples directores. En cantidad de capítulos, no llegarían a dirigir una temporada completa.