Práctica 1: Analisis de Reviews sobre Amazon

Aprendizaje Estadístico

Andrés Campos Cuiña

01/10/2021

Índice

1. Carga de las librerías necesarias

En primer lugar, cargamos las librerías necesarias:

```
library(tidyverse) # General-purpose data wrangling
```

Definimos el directorio de trabajo:

```
setwd("c:/Users/Andres/Google Drive/USC/MaBD/Aprendizaje Estadistico/Practicas/Practica1")
```

2. Carga de los datos con los que se trabajará

Ahora cargamos los datos en R mediante el uso de la función `read_tsv`:

```
data <- read_tsv("data/amazon.tsv")
data <- as.data.frame(data)
```

3. Ejercicio 1

¿Cuántos comentarios han sido registrados? ¿Cómo es la distribución del número de comentarios a lo largo de los años?

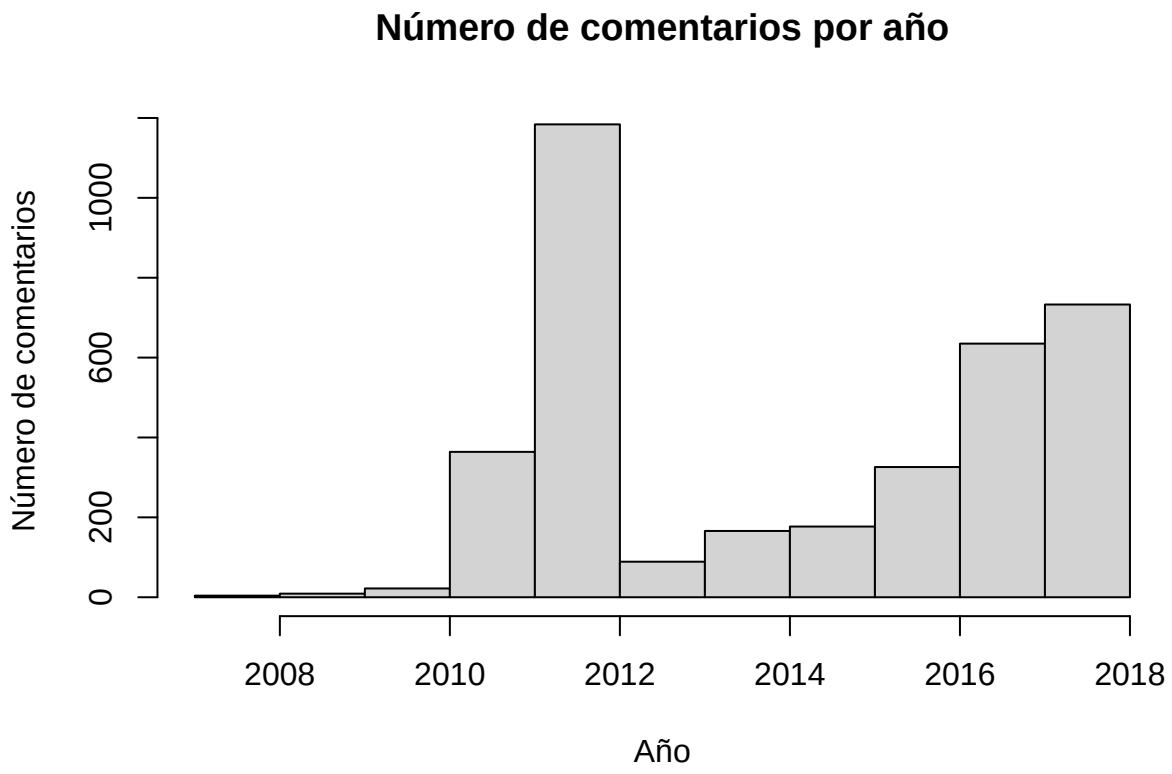
Número de comentarios:

```
n_reviews <- nrow(data)
print(paste("Nº de comentarios:", n_reviews, sep = " "))
```

```
## [1] "Nº de comentarios: 3709"
```

Distribución del número de comentarios a lo largo de los años:

```
reviews_by_year <- c(str_split_fixed(data$date, '-', n = 2)[, 1])
reviews_by_year <- strtoi(reviews_by_year, base = 0L)

hist(reviews_by_year,
     main = "Número de comentarios por año",
     xlab = "Año",
     ylab = "Número de comentarios"
)
```



Como se puede observar en la gráfica superior los comentarios empiezan a aparecer en el año 2007. De 2007 a 2012 el número de comentarios aumenta cada año, llegando a su maximo en el año 2012. Por algún motivo el número de comentarios cae drásticamente del año 2012 al año 2013. Después de esta brusca caída, el número de comentarios por año vuelve a aumentar año a año hasta el 2018.

4. Ejercicio 2

¿Cómo resumirías la valoración de Amazon según los usuarios? ¿Cómo ha evolucionado la valoración media por año?

Para resumir la valoración media de Amazon por parte de los usuarios podríamos tomar la media de todos las comentarios presentes en el fichero de datos:

```
rating_avg <- mean(data$rating)
print(paste("Media de todos los comentarios:", rating_avg, sep = " "))
```

```
## [1] "Media de todos los comentarios: 4.19870585063359"
```

Como se puede ver la media de todos los comentarios presentes en el fichero de datos es, aproximadamente de 4.2 sobre 5, por lo que podemos concluir que la valoración de Amazon es positiva.

Podemos utilizar la función `summary()` para ver un poco más en detalle la información sobre las valoraciones:

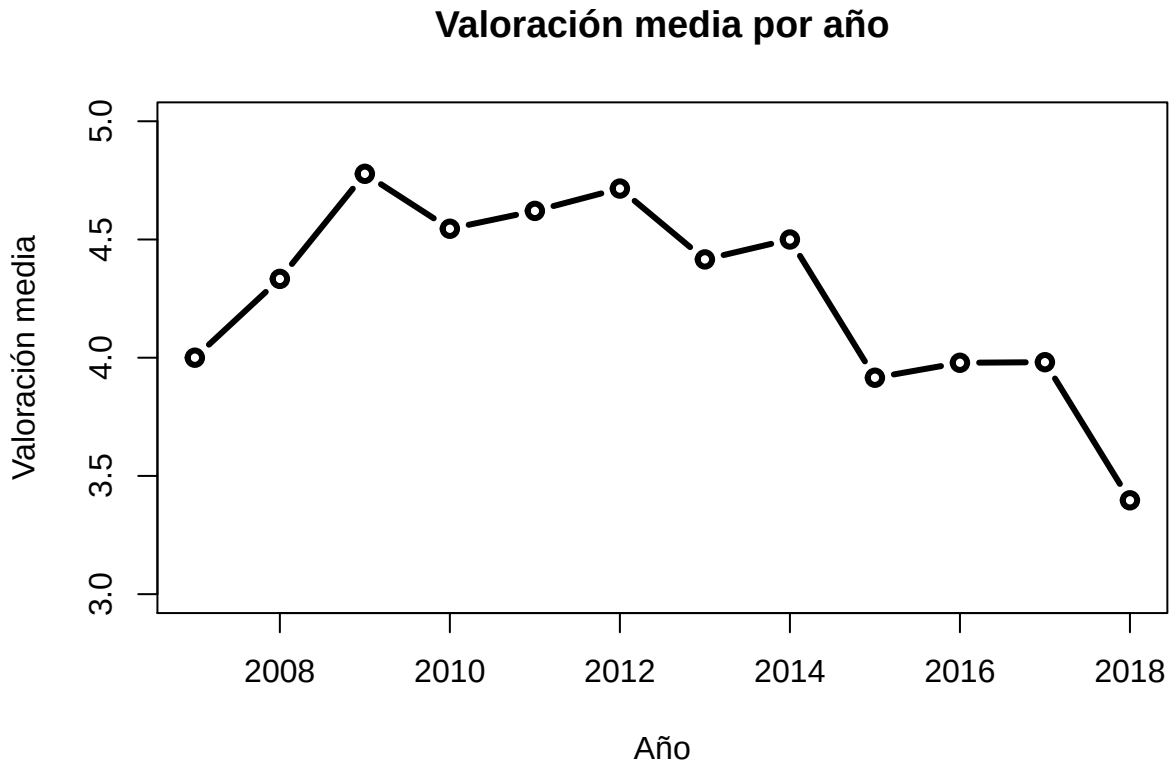
```
summary(data$rating)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   4.000   5.000   4.199   5.000   5.000
```

Como se puede observar, el primer cuartil (Q1) es igual a 4, por lo que sólo el 25 % de las valoraciones son inferiores a esta puntuación, que es bastante alta teniendo en que se usa una escala del 1 al 5.

Cómo ha evolucionado la valoración media por año:

```
ratings <- data[c("date", "rating")]
ratings <-
  mutate(ratings, date = strtoi(str_split_fixed(date, '-', n = 2)[, 1], base = 0L))

rating_avg_by_year <-
  ratings %>% group_by(date) %>% summarise(mean = mean(rating))

plot(
  rating_avg_by_year,
  type = "b",
  lwd = "3",
  main = "Valoración media por año",
  xlab = "Año",
  ylab = "Valoración media",
  ylim = c(3, 5)
)
```



En la gráfica superior se puede observar como desde el 2007 al 2009 las valoraciones medias (partiendo de una valoración media igual a 4 en el año 2007) van creciendo, por lo que la satisfacción de los clientes con Amazon durante estos años fue en aumento. Del 2009 al 2012 la valoración media se mantiene relativamente estable entorno al 4.7, sin embargo, a partir del año 2012 esta comienza a descender hasta llegar a una valoración media igual a 3.5 en el año 2018. Esto probablemente sea debido a un peor servicio por parte de Amazon al aumentar rápidamente el numero de clientes durante estos años.

5. Ejercicio 3

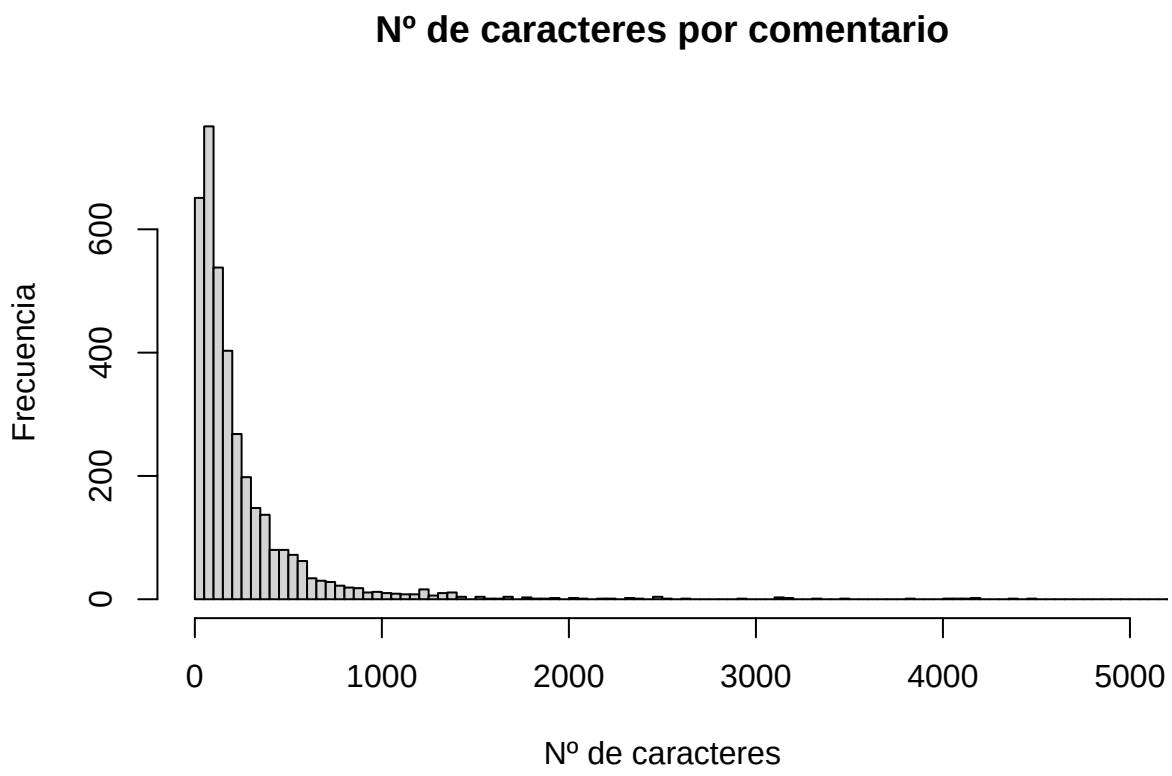
Representa y comenta la distribución del número de caracteres de los comentarios

En primer lugar calculamos mediante la función `nchar` el número de caracteres por comentario para cada uno de los comentarios presentes en nuestro conjunto de datos:

```
# Para cada comentario obtenemos el número de caracteres
# y lo guardamos en una nueva lista
n_characters_by_review <- c()

for (review in data$review) {
  n_char <- nchar(review, type="chars")
  n_characters_by_review <- append(n_characters_by_review, n_char)
}

data["n_characters"] <- n_characters_by_review
```

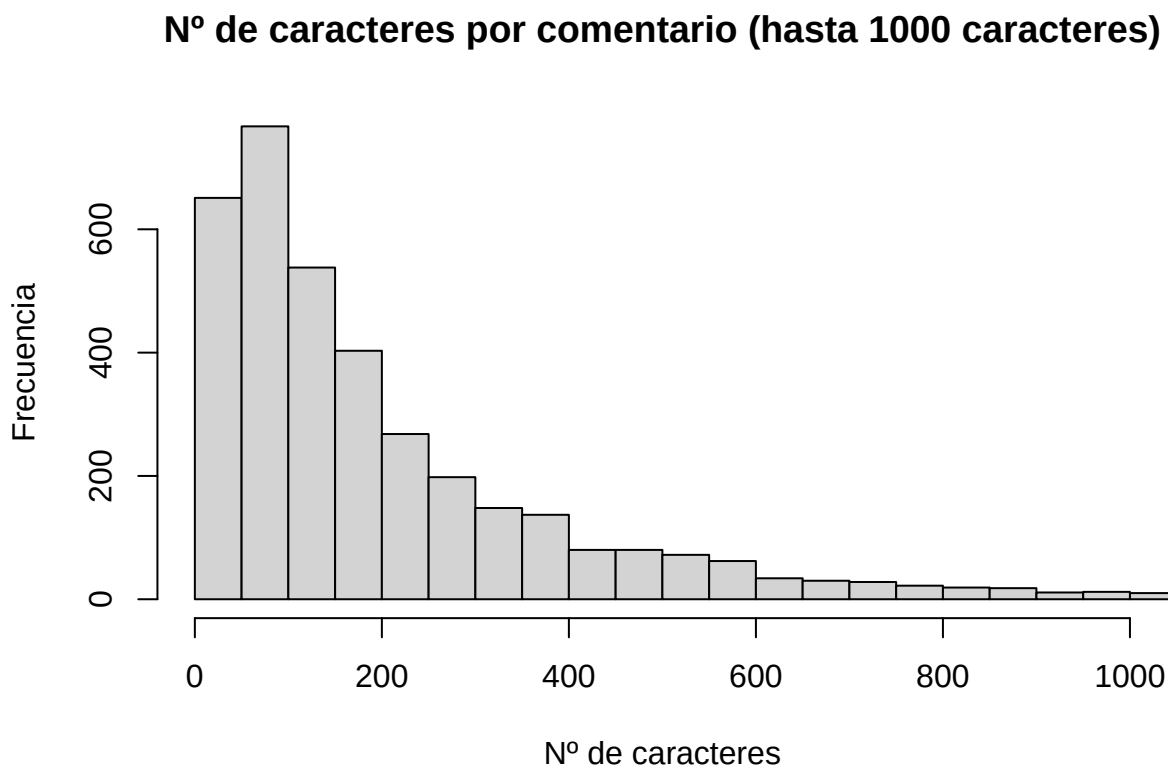
Ahora representamos mediante un histograma (usando la función `hist`) la distribución del número de caracteres de los comentarios:

```
hist(
  data$n_characters,
  breaks=200,
  xlim=c(0,5000),
  main="Nº de caracteres por comentario",
  xlab = "Nº de caracteres",
  ylab = "Frecuencia",
)
```



Como se puede observar la gran mayoría de los comentarios cuentan con un número de comentarios menor que mil, por lo que veremos la distribución en el intervalo de 0 a 1000 [0, 1000] para más detalle:

```
hist(
  data$n_characters,
  breaks=200,
  xlim=c(0,1000),
  main="Nº de caracteres por comentario (hasta 1000 caracteres)",
  xlab = "Nº de caracteres",
  ylab = "Frecuencia",
)
```



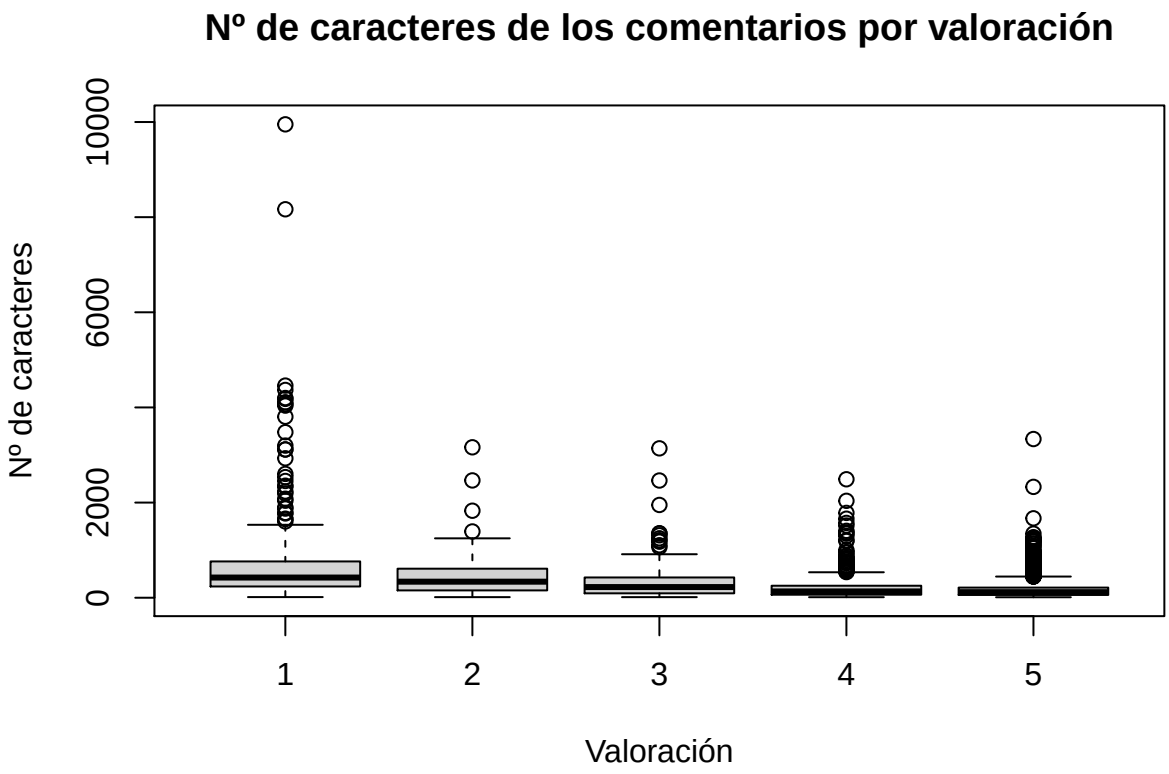
En la gráfica superior podemos observar como el número de caracteres por comentario sigue una distribución que podría ser aproximada por una distribución log-normal. El intervalo con más comentarios es el de 50 a 100 caracteres, estando la mayor parte de los comentarios por debajo de los 200 caracteres. Esto nos dice que la mayor parte de la gente deja comentarios relativamente cortos, sin embargo, también podemos observar como aún así hay comentarios de una extensión mucho mayor, llegando hasta los 4000 caracteres incluso.

6. Ejercicio 4

Representa mediante un diagrama de cajas la distribución del número de caracteres de los comentarios en función de la valoración (de 1 a 5). ¿Qué observas?

En primer lugar representamos mediante diagramas de caja el número de caracteres de los comentarios en función de su valoración:

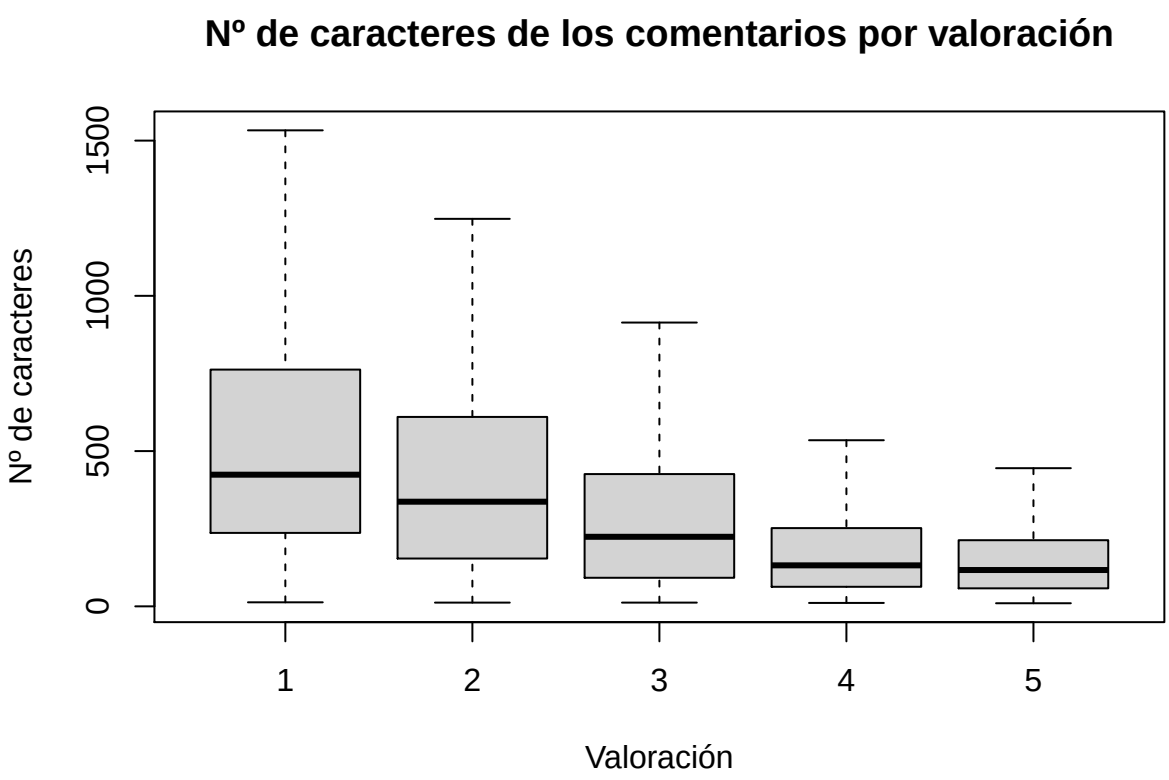
```
boxplot(
  data$n_characters ~ data$rating,
  main = "Nº de caracteres de los comentarios por valoración",
  xlab = "Valoración",
  ylab = "Nº de caracteres",
)
```



Como se puede observar para todas la valoraciones el número de `outliers` es muy elevado (siendo una excepción los comentarios con una valoración de 2). Este elevado número de `outliers` nos dice que la longitud de los comentarios es muy variada para todas las valoraciones. Es destacable el caso de los comentarios de una valoración igual a 1. En este caso hay dos `outliers` que destacan mucho por su gran número de caracteres, teniendo uno de ellos sobre 8000 caracteres y el otro sobre 10000. De estos dos comentarios con tan baja puntuación pero tan larga extensión podemos deducir el claro descontento con la empresa de los dos usuarios que dejaron esos comentarios.

Para pode observar mejor la distribución del número de caracteres por comentario en función de la valoración, representaremos de nuevo sin mostrar los `outliers`:

```
boxplot(
  data$n_characters ~ data$rating,
  main = "Nº de caracteres de los comentarios por valoración",
  xlab = "Valoración",
  ylab = "Nº de caracteres",
  outline=FALSE
)
```



Ahora vemos más claramente como a medida que la valoración aumenta, el número de caracteres de los comentarios se reduce, es decir, que la gente escribe más cuando quiere expresar una opinión negativa que una positiva. También podemos ver como a medida que aumenta la valoración, el número de caracteres por comentario se acerca más a la media para esa valoración, es decir, la longitud de los comentarios es más similar entre si para los comentarios de esa valoración. Podemos ver como para los comentarios de una valoración igual a 1 la longitud media de los comentarios es aproximadamente de unos 500 caracteres, mientras que para los comentarios de una valoración igual a 5 la longitud media se encuentra sobre los 150 caracteres.

7. Ejercicio 5

Completa el análisis descriptivo de los datos con información que consideres relevante

Para completar el análisis descriptivo de los datos podríamos ampliar un poco más la información que tenemos acerca de los usuarios, para esto estudiaremos el número de comentarios por usuario (con el fin de si los usuarios son variados o si un gran número de los comentarios son hechos por un pequeño grupo de usuarios, a los que se les podría dar más importancia) y la valoración media de los comentarios dejados por cada uno de los usuarios (con el fin de estudiar si algunos usuarios dejan muy malas o muy buenas valoraciones siempre, lo que podría significar que sus valoraciones no son del todo objetivas).

Empezaremos por calcular el número de comentarios por usuario y mostraremos a aquellos usuarios que más comentarios han dejado:

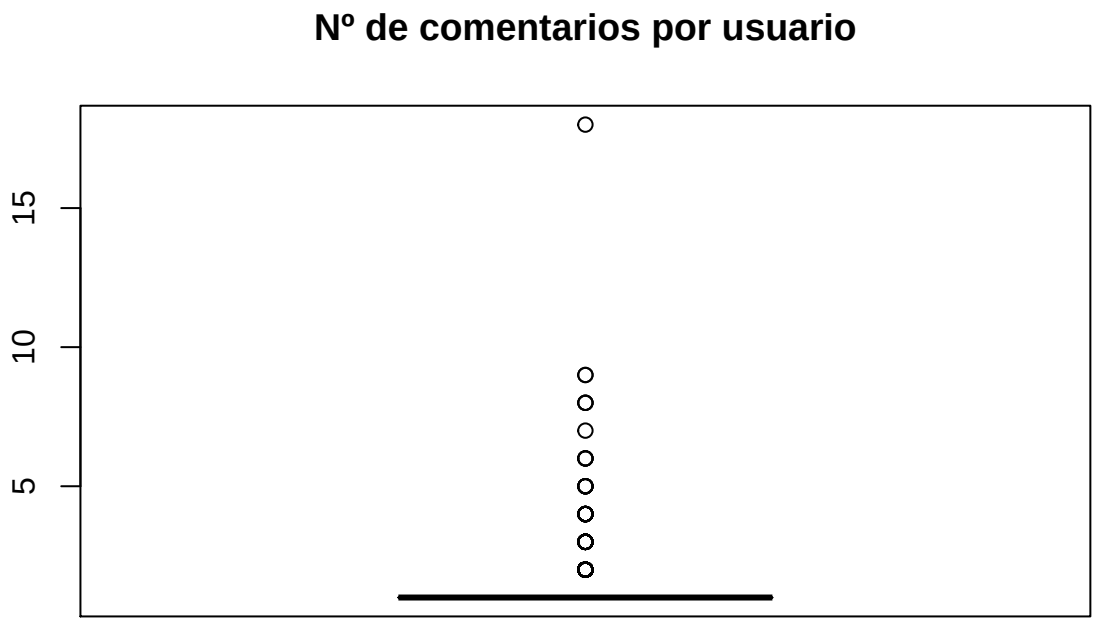
```
n_reviews_by_reviewer <-
  data %>% count(reviewer) %>% arrange(desc(n))

# Imprimimos los 5 usuarios con más comentarios
print(head(n_reviews_by_reviewer, n = 5))
```

```
##   reviewer  n
## 1 Customer 18
## 2     John  9
## 3     Mike  9
## 4    David  8
## 5  Michael  8
```

Como se puede observar hay ciertos usuarios que tienen más comentarios que otros, suponiendo que los nombres de usuario son único y no repetibles, estos sería los usuarios más activos en la web de reviews.

Ahora representaremos mediante un diagrama de cajas la distribución del número de comentarios por usuario:

```
boxplot(
  n_reviews_by_reviewer$n,
  main = "Nº de comentarios por usuario"
)
```

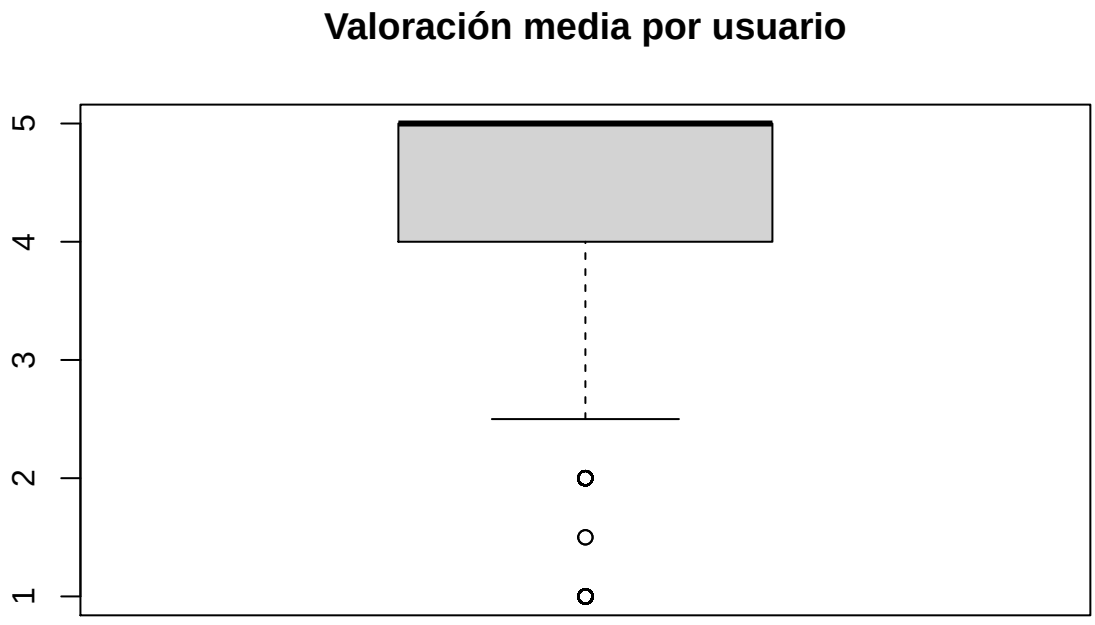


Como se puede ver sí que hay usuarios con más de un comentario, sin embargo, la gran mayoría de los usuarios sólo han dejado 1 comentario. Esto nos puede llevar a destacar los comentarios dejados por los usuarios recurrentes o a darle más importancia a los mismos.

Por último, estudiaremos la valoración media por usuario:

```
rating_avg_by_reviewer <-
  data %>% group_by(reviewer) %>% summarise(
    mean = mean(rating)) %>% mutate_if(is.numeric, ~ round(., 1))

boxplot(
  rating_avg_by_reviewer$mean,
```

```
  main = "Valoración media por usuario"
)
```



Como se puede observar la mayor parte de los usuarios deja valoraciones medias entre 4 y 5. No obstante, hay ciertos `outliers` que dejan valoraciones medias inferiores a 2. Amazon podría estudiar los comentarios de estos usuarios para tratar de resolver sus críticas y tratar de mejorar el servicio.