Práctica 2 - Tipología y Ciclo de Vida de los Datos

Esteban Braganza Cajas y Ana Álvarez Sánchez

2024-05-22

1. Descripción del dataset

¿Por qué es importante y qué pregunta/problema pretende responder? Resume brevemente las variables que lo forman y su tamaño.

PREGUNTA/PROBLEMA A RESPONDER:

```
data <- read.csv("../data/images_db.csv")
```

El fichero de datos contiene 7067 registros y 18 variables.

El dataset está compuesto por los siguientes campos:

- `search_topic`: la imagen es resultado de la búsqueda por este tema
- `page_num`: la imagen aparece en este número de página de la búsqueda
- `image_page`: enlace a la página con la información de la imagen
- `image_url`: enlace a la imagen
- `image_title`: título de la imagen
- `image_author`: autor/a de la imagen
- `image_favs`: número de veces que le han dado a "me gusta" en la imagen
- `image_com`: número de comentarios que tiene la imagen
- `image_views`: número de vistas a la imagen
- `private_collections`: número de veces que ha sido incluida en una colección privada
- `tags`: etiquetas que se le han asignado a la imagen para facilitar su descubrimiento
- `location`: país o localización geográfica, si el autor la quiere identificar description: campo de texto abierto creado por el autor, que acompaña a la imagen. Puede incluir detalles técnicos o enlaces a las redes sociales del autor/a.
- `description`; descripción de la imagen que hace el autor.
- `image_px`: dimensiones de la imagen, en pixeles
- `image_size`: peso de la imagen en MB.
- `published_date`: fecha de publicación de la imagen.
- `last_comment`: último comentario añadido a la imagen.
- `license`: licencia de la imagen

En cuanto a la descripción de cada una de estas variables:

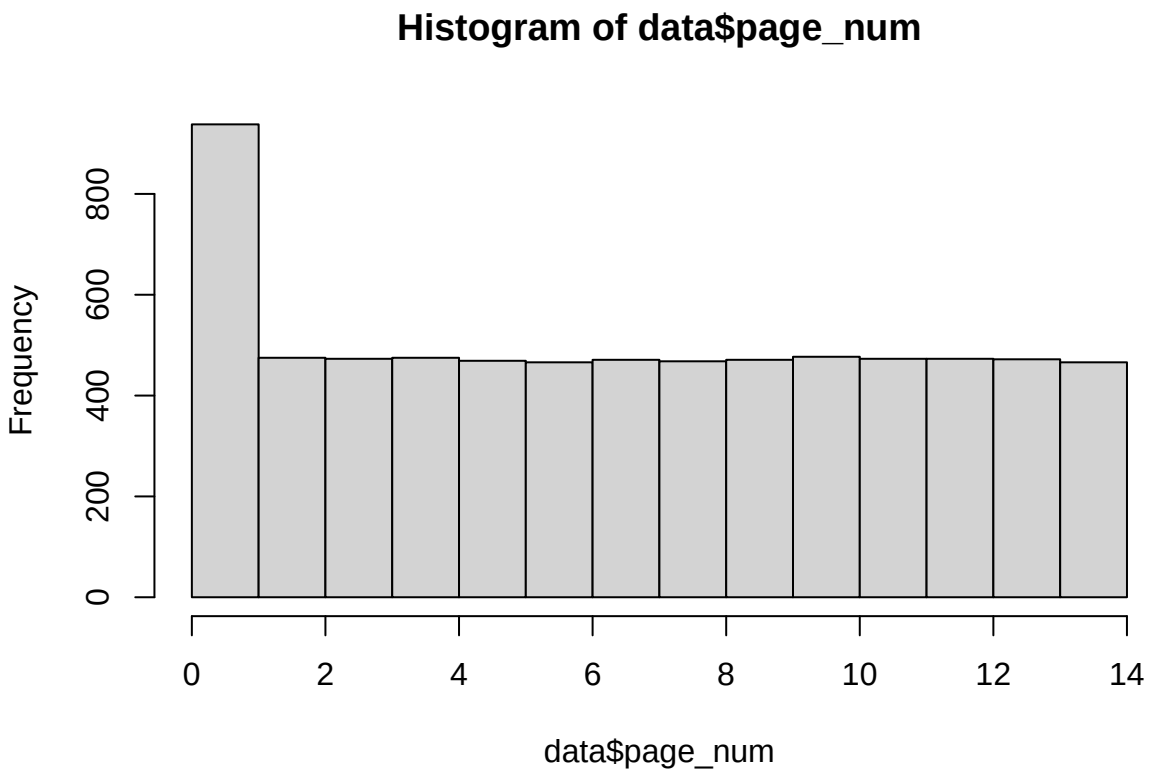
- La variable `search_topic` incluye 20 posibilidades, que son: Fantasy art, Science fiction art, Anime and manga art, Fan art (for specific fandoms), Digital paintings, Traditional drawings, Character designs, Creature concepts, Landscape art, Abstract art, Surrealism, Steampunk art, Cyberpunk art, Gothic art, Horror art, Cosplay photography, Pixel art, Concept art, Comics and graphic novels, Street art and graffiti. Las frecuencias de cada una de estas categorías en el conjunto de datos es esta:

```
knitr::kable(table(data$search_topic))
```

Var1	Freq
Abstract art	352
Anime and manga art	348
Character designs	343
Comics and graphic novels	351
Concept art	358
Cosplay photography	359
Creature concepts	353
Cyberpunk art	358
Digital paintings	347
Fan art (for specific fandoms)	346
Fantasy art	359
Gothic art	352
Horror art	358
Landscape art	360
Pixel art	355
Science fiction art	351
Steampunk art	354
Street art and graffiti	357
Surrealism	348
Traditional drawings	358

- La variable `page_num` incluye 15 posibilidades, que están en el rango 0, 14, porque en la descarga se han elegido las primeras 15 páginas de cada topic buscado.

```
hist(data$page_num)
```



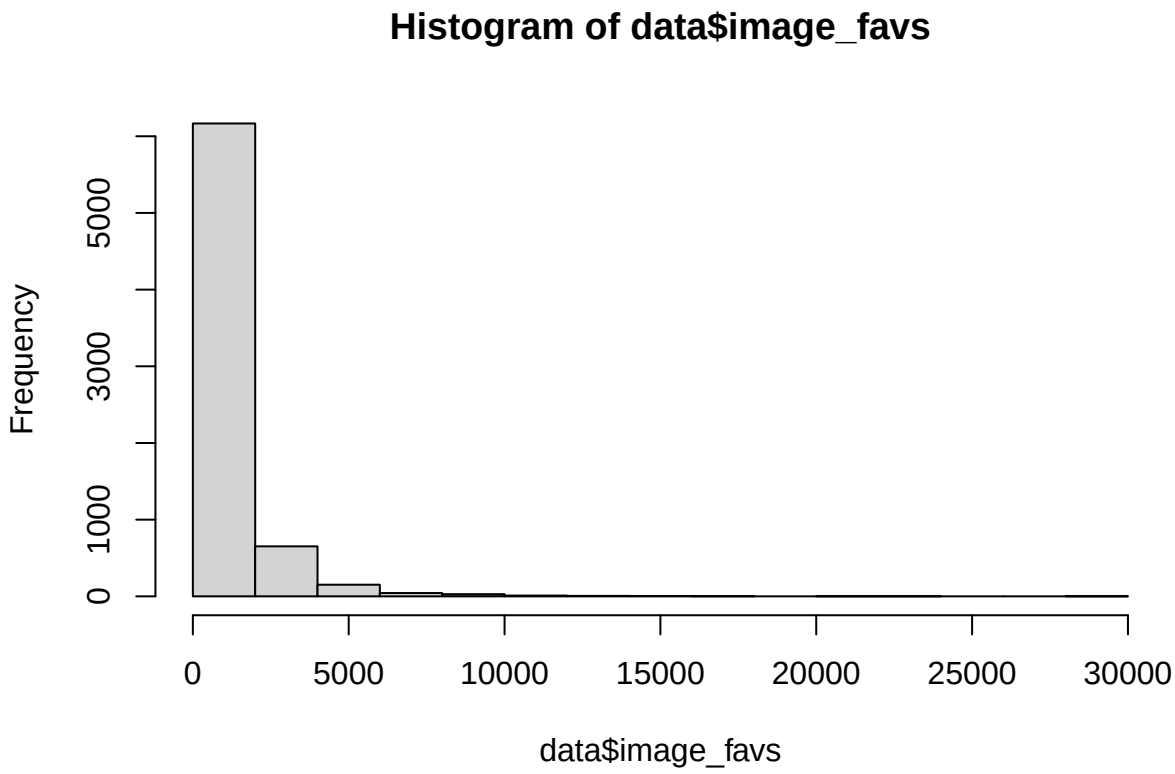
- Las variables `image_page`, `image_url`, `image_title`, `description` y `last_comment` son variables de tipo texto.
- La variable `image_author` incluye 2903 posibilidades. El autor/a con más registros tiene 142 obras. 2047 tienen una sola obra en el dataset. La media de imágenes por autor es de 2.43.

```
frecuencias <- table(data$image_author)
df_frecuencias <- as.data.frame(frecuencias)
df_frecuencias_ordenado <- df_frecuencias[order(-df_frecuencias$Freq), ]

knitr::kable(head(df_frecuencias_ordenado, 10))
```

	Var1	Freq
1489	Leonidafremov	142
334	BisBiswas	140
1159	IrenHorrors	131
1680	MichaelAdamidisArt	102
1243	JJcanvas	76
1294	Julian-Faylona	76
1828	Nele-Diel	74
504	Cloister	63
255	ARVEN92	61
159	AndrejZT	53

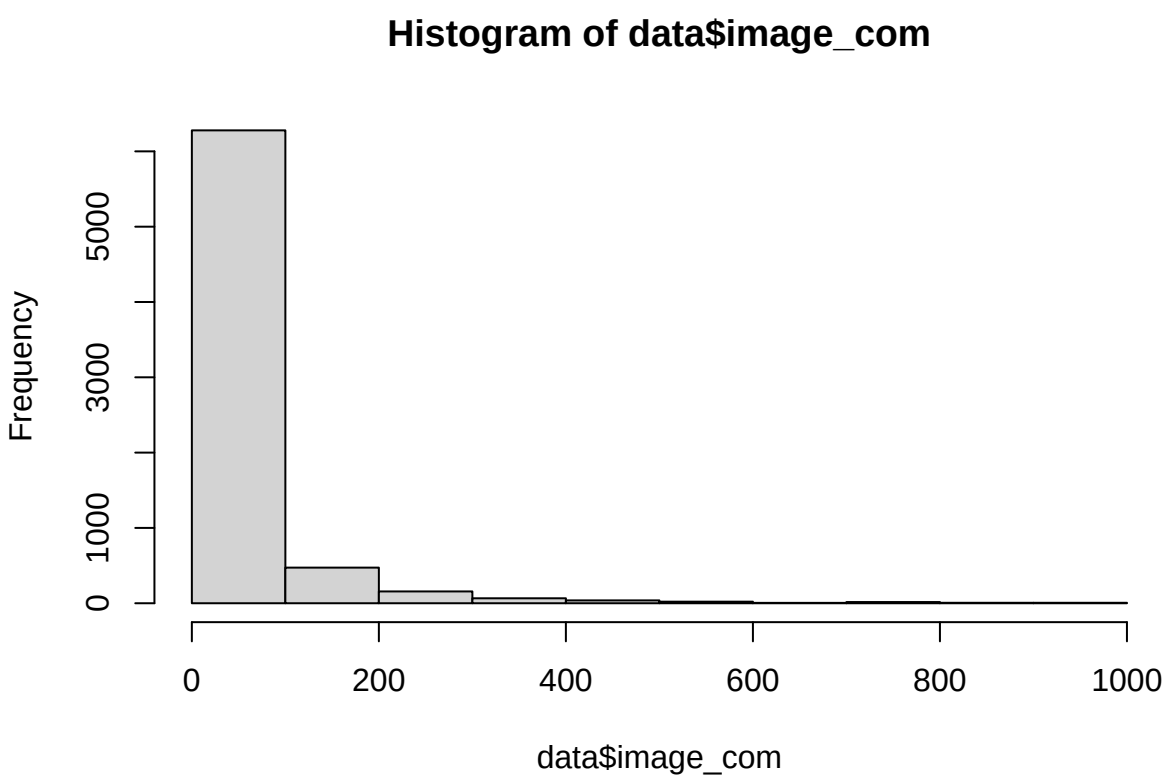
- La variable `image_favs` es numérica, con un valor mínimo de 0, un valor máximo de 28300, una media de 967.7, con esta distribución:

```
hist(data$image_favs)
```



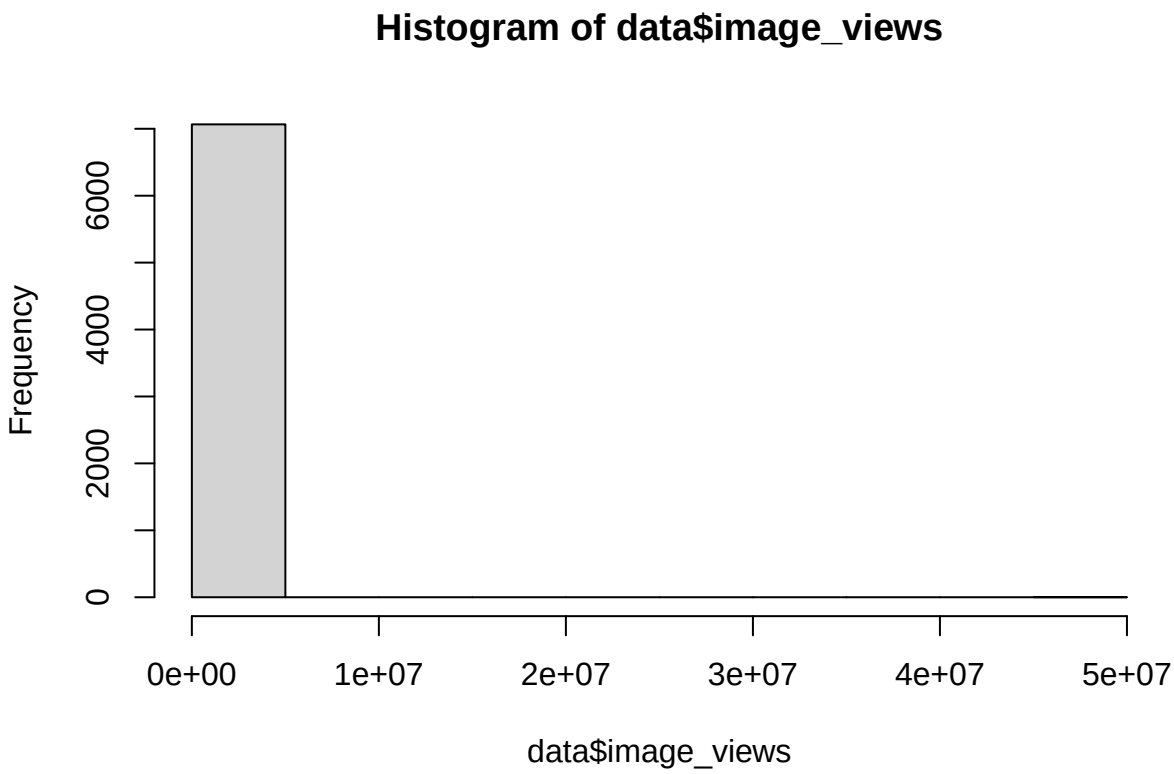
- La variable `image_com` es numérica, con un valor mínimo de 0, un valor máximo de 986, una media de 47.93, con esta distribución:

```
hist(data$image_com)
```



- La variable `image_views` es numérica, con un valor mínimo de 1, un valor máximo de 48400000, una media de 9.926775×10^4, una mediana de 20900, con esta distribución:

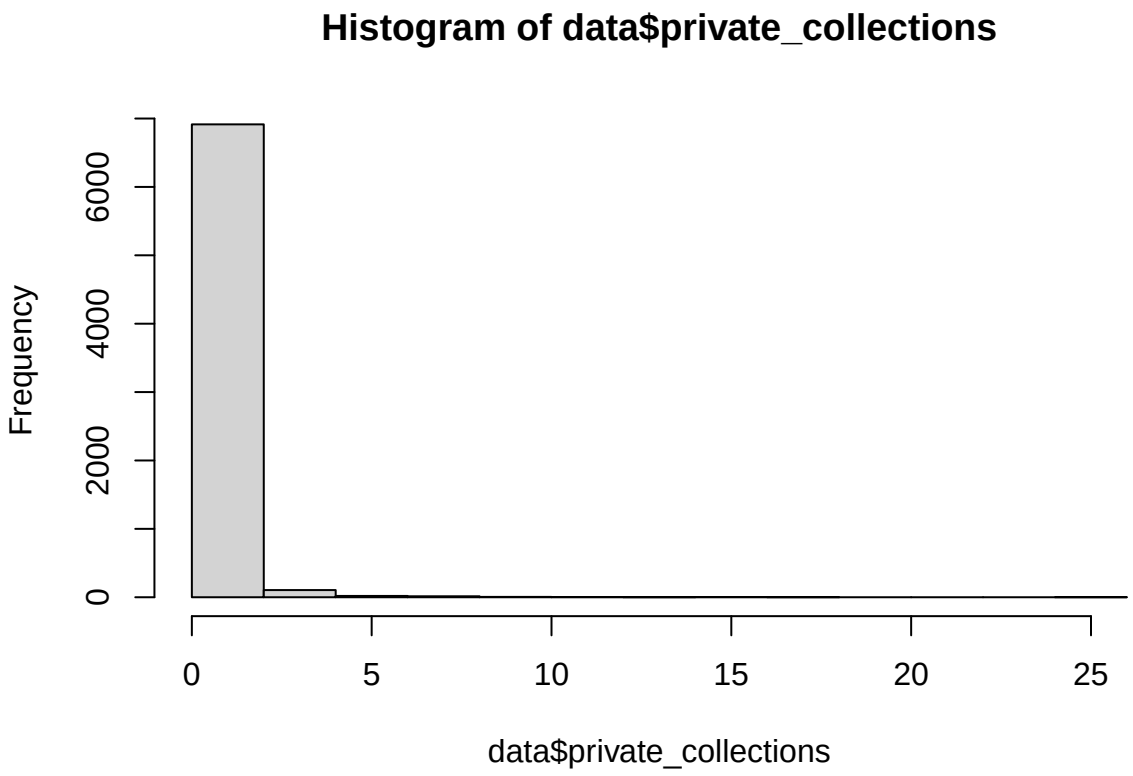
```r
hist(data$image_views)
```



Esto indica la presencia de outliers en esta variable.

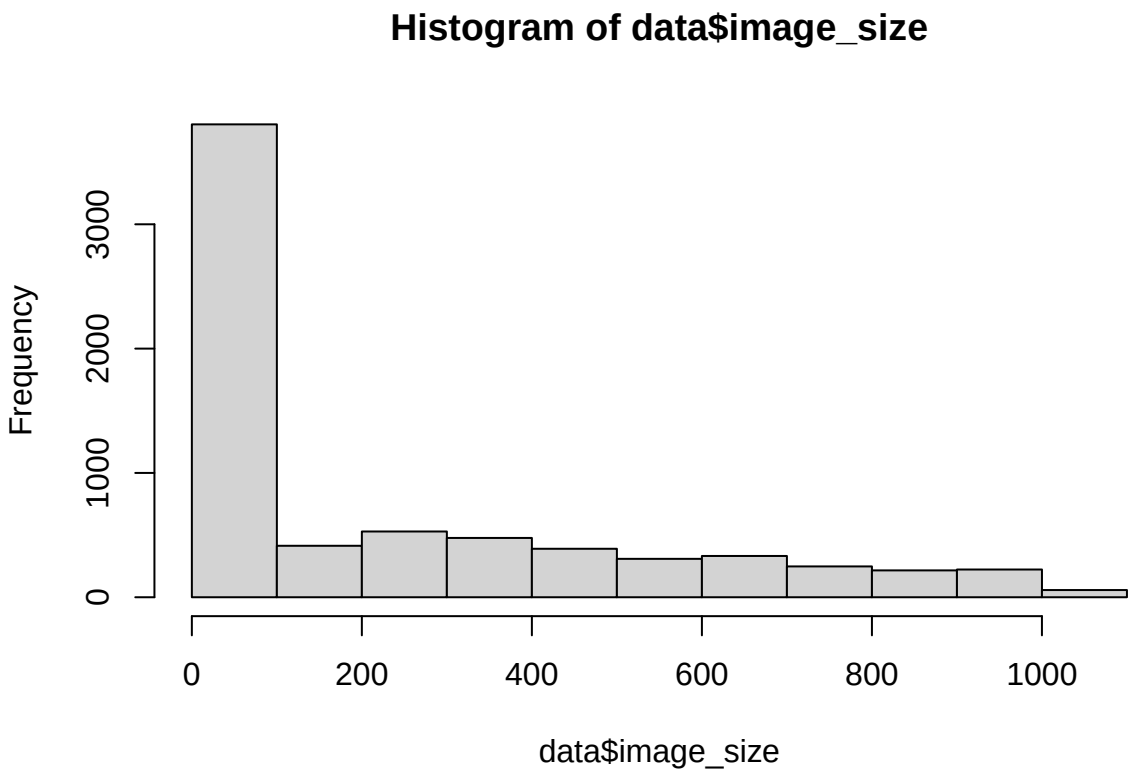
- La variable `private_collections` es numérica, con un valor mínimo de 0, un valor máximo de 26, una media de 0.2, con esta distribución:

```
hist(data$private_collections)
```



- La variable `tags` -> qué hacer con ella????
- La variable `image_size` es numérica, con un valor mínimo de NA, un valor máximo de NA, una media de NA, con esta distribución:

```
hist(data$image_size)
```



Se limpian variables `location` (candidata a ser eliminada por número de NA), `published_date` (conversión a tipo Date), `image_license`, unificando las licencias copyright y convirtiéndola a categórica, `image_px` (unificamos formato en ancho x largo y extraemos superficie de la imagen, el resto de valores los convertimos a NA). Se desarrolla en el apartado de limpieza de los datos.

2. Integración y selección de los datos de interés a analizar

Puede ser el resultado de adicionar diferentes datasets o una subselección útil de los datos originales, en base al objetivo que se quiera conseguir. Si se decide trabajar con una selección de los datos, es muy importante que esta esté debidamente justificada. Además, se recomienda mostrar un resumen de los datos que permita ver a simple vista las diferentes variables y sus rangos de valores.

PROPUESTA: eliminación de las variables de texto libre, como `description`, `image_url`, `image_page`, `last_comment`. Eliminación de las variables con gran número de NA (`location`). Mantener una de las dos variables de dimensiones: `size` o `superficie`, eliminar la otra.

```
summary(data)
```

```
##  search_topic          page_num   image_page         image_url
##  Length:7067        Min.   : 0   Length:7067        Length:7067
##  Class :character   1st Qu.: 3   Class :character   Class :character
##  Mode  :character   Median : 7   Mode  :character   Mode  :character
##                     Mean   : 7
##                     3rd Qu.:11
##                     Max.   :14
```

```
##
##  image_title        image_author         image_favs        image_com
##  Length:7067        Length:7067        Min.   :    0.0   Min.   :  0.00
##  Class :character   Class :character   1st Qu.:  113.5   1st Qu.:  7.00
##  Mode  :character   Mode  :character   Median :  508.0   Median : 22.00
##                                        Mean   :  967.7   Mean   : 47.93
##                                        3rd Qu.: 1200.0   3rd Qu.: 52.00
##                                        Max.   :28300.0   Max.   :986.00
##
##   image_views       private_collections     tags              location
##  Min.   :       1   Min.   : 0.0000     Length:7067        Length:7067
##  1st Qu.:    4200   1st Qu.: 0.0000     Class :character   Class :character
##  Median :   20900   Median : 0.0000     Mode  :character   Mode  :character
##  Mean   :   99268   Mean   : 0.1956
##  3rd Qu.:   67550   3rd Qu.: 0.0000
##  Max.   :48400000   Max.   :26.0000
##
##  description          image_px           image_size      published_date
##  Length:7067        Length:7067        Min.   :   1.00   Length:7067
##  Class :character   Class :character   1st Qu.:   3.01   Class :character
##  Mode  :character   Mode  :character   Median :  24.68   Mode  :character
##                                        Mean   : 227.92
##                                        3rd Qu.: 404.67
##                                        Max.   :1023.92
##                                        NA's   :67
##  last_comment       image_license
##  Length:7067        Length:7067
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

3. Limpieza de los datos

3.1. ¿Los datos contienen ceros, elementos vacíos u otros valores numéricos que indiquen la pérdida de datos? Gestiona cada uno de estos casos utilizando el método de imputación que consideres más adecuado. 3.2. Identifica y gestiona adecuadamente el tipo de dato de cada atributo (p.ej. conversión de variables categóricas en factor). 3.3. Identifica y gestiona los valores extremos. 3.4. Justifica la necesidad de otros métodos de limpieza para este dataset en particular y, de ser necesario, aplícalos.

- Conversión a categórica de la variable `search_topic`:

```
data$search_topic <- as.factor(data$search_topic)
```

- La variable `location` tiene un alto porcentaje de valores vacíos (" "), que constituyen el 96.79% de los registros. Convertimos esos valores en NA para su mejor identificación.

```
data$location[data$location == ""] <- NA
```

- La variable `published_date` es de tipo character, la convertimos a tipo Date.

```
# Extraer solo la parte de la fecha (sin la hora)
data$published_date <- substr(data$published_date, 1, 10)

# Convertir a formato de fecha
data$published_date <- as.Date(data$published_date)
```

- La variable `image_license`, contiene muchos valores vacíos (" "). Empezamos convirtiendo esos en NA.

```
data$image_license[data$image_license == ""] <- NA
```

Todas las que tengan el símbolo "©" las sustituimos por la palabra Copyright:

```
# Identificar las filas que contienen "©"
filas_con_copyright <- grep("©", data$image_license)

# Sustituir toda la celda por "Copyright" para las filas identificadas
data$image_license[filas_con_copyright] <- "Copyright"
```

```
knitr::kable(table(data$image_license))
```

Var1	Freq
Copyright	6276
Creative Commons Attribution 3.0 License	40
Creative Commons Attribution-No Derivative Works 3.0 License	24
Creative Commons Attribution-Noncommercial 3.0 License	8
Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 License	545
Creative Commons Attribution-Noncommercial-Share Alike 3.0 License	19
Creative Commons Attribution-Share Alike 3.0 License	20

La convertimos en categórica:

```
data$image_license <- as.factor(data$image_license)

str(data$image_license)
```

```
##  Factor w/ 7 levels "Copyright","Creative Commons Attribution 3.0 License",..: 1 1 1 1 1 5 1 1 1 1 .
```

- La variable `image_px` es de tipo character e indica el tamaño de la imagen en píxeles, en formato "ancho x alto". Para poder utilizarla en nuestro análisis como una variable numérica. Antes de convertir los dos valores en numéricos, comprobamos que todos los datos aparecen en ese formato "ancho x alto":

```
# Obtener los índices de las filas que no contienen "x"
filas_sin_x <- grep("x", data$image_px, invert = TRUE)

# Ver las filas que no contienen "x"
nrow(data[filas_sin_x, "image_px", drop = FALSE])
```

```
## [1] 67
```

Estos datos podrían representar el número total de píxeles en la imagen, en lugar de especificar el ancho y el alto por separado.

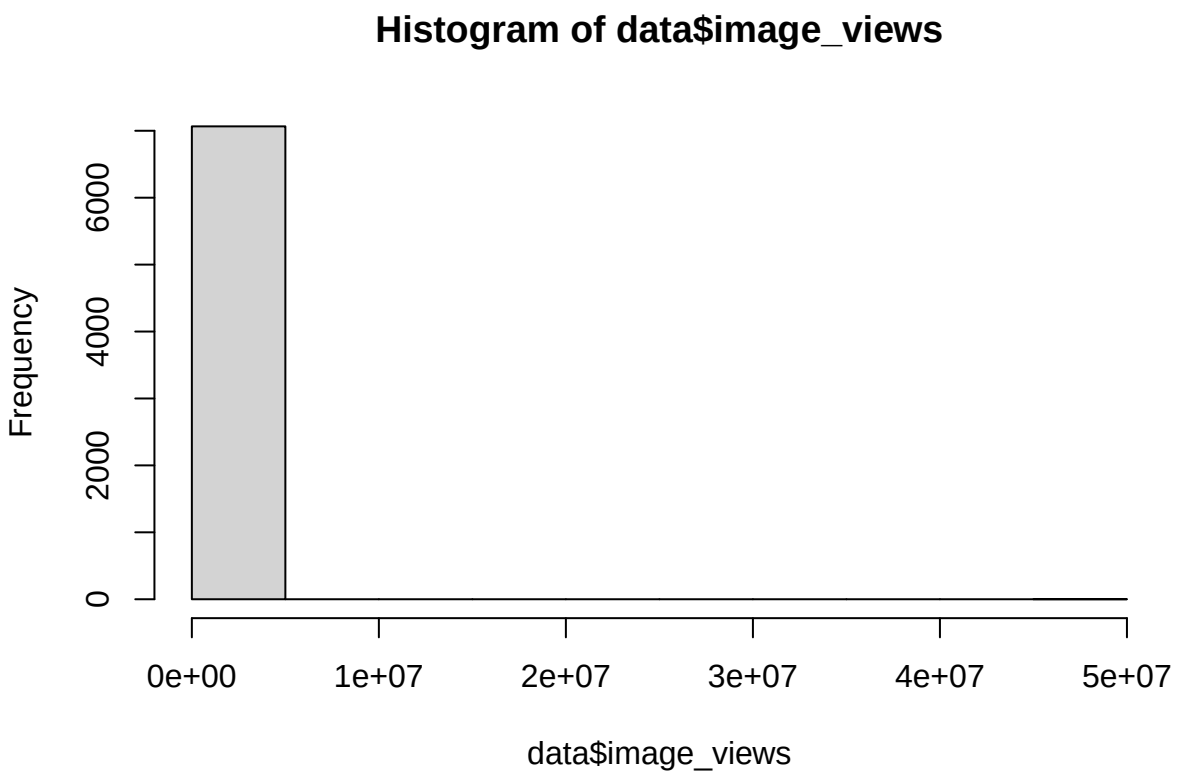
Opción: ignorar estos valores, convirtiéndolos en NA, porque no podemos saber si las imágenes son cuadradas (mismo ancho y alto)

```
data$image_px[filas_sin_x] <- NA
```

```
# Filtrar las filas que no tienen valores NA en la columna image_px
sin_na <- complete.cases(data$image_px)

# Dividir la cadena en dos partes (ancho y alto) solo para las filas sin NA
dimensiones <- strsplit(data$image_px[sin_na], "x")

# Convertir las partes a valores numéricos
ancho <- as.numeric(sapply(dimensiones, "[[", 1))
alto <- as.numeric(sapply(dimensiones, "[[", 2))

# Crear nuevas columnas en el data frame para el ancho y el alto y la superficie total
# data$ancho[sin_na] <- ancho
# data$alto[sin_na] <- alto
data$superficie[sin_na] <- ancho * alto
```

- La variable `image_views` es numérica, con un valor mínimo de 1, un valor máximo de 48400000, una media de 9.926775×10^4, una mediana de 20900, con esta distribución:

```
hist(data$image_views)
```



Esto indica la presencia de outliers en esta variable.

Convertimos todos los valores en el dataset que son texto vacío (“ “) en NA, por si han quedado en alguna otra variable, puesto que se trata de valores ausentes.

```
# Convertir todos los valores vacíos ("") en el dataset a NA
data <- data.frame(lapply(data, function(x) {
  x[x == ""] <- NA
  return(x)
}), stringsAsFactors = FALSE)
```

Valores ausentes de cada variable:

```
colSums(is.na(data))
```

```
##        search_topic            page_num          image_page           image_url
##                   0                   0                   0                   0
##         image_title        image_author          image_favs           image_com
##                   0                   0                   0                   0
##         image_views private_collections                tags            location
##                   0                   0                   0                6840
##         description           image_px          image_size      published_date
##                 453                  67                  67                   0
##        last_comment       image_license          superficie
##                2734                 135                  67
```

4. Análisis de los datos

4.1. Aplica un modelo supervisado y uno no supervisado a los datos y comenta los resultados obtenidos.
4.2. Aplica una prueba por contraste de hipótesis. Ten en cuenta que algunas de estas pruebas requieren verificar previamente la normalidad y homocedasticidad de los datos.

5. Representación de los resultados a partir de tablas y gráficas

Este apartado se puede responder a lo largo de la práctica, sin necesidad de concentrar todas las representaciones en este apartado.

Se debe representar tanto el contenido del dataset para observar las proporciones y distribuciones de las diferentes variables una vez aplicada la etapa de limpieza, como los resultados obtenidos tras la etapa de análisis.

6. Resolución del problema

A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?