Práctica 2: Limpieza y validación de los datos

Autor: Alberto Giménez Aragón

Enero 2021

Contents

1 Descripción del dataset

Para la realización de esta práctica he escogido el dataset que generamos en la práctica 1. Este dataset recoge una gran variedad de datos sobre los distintos hoteles repartidos a lo largo de la ciudad de Barcelona y alrededores, los cuales fueron obtenidos mediante *web scraping* a fecha de 2 de noviembre de 2020.

A continuación, se detallan todos los atributos de los que dispone el dataset.

- **Name**: Nombre del Hotel
- **Stars**: Estrellas del hotel, entre 1 y 5 estrellas
- **Score**: Calificación del hotel según los usuarios.
- **Score Location**: Calificación de la ubicación del hotel según los usuarios (de 1.0 a 5.0).
- **Score Cleaning**: Calificación de la limpieza del hotel según los usuarios (de 1.0 a 5.0).
- **Score Service**: Calificación del servicio del hotel según los usuarios (de 1.0 a 5.0).
- **Score Value for Money**: Calificación de la relación calidad/precio del hotel según los usuarios (de 1.0 a 5.0).
- **Price**: Precio actual por noche del hotel
- **Price Range**: Rango de precios por noche en el que se encuentra el hotel
- **Ranking**: Ranking Tripadvisor del hotel respecto a otro hoteles de la ciudad
- **Number opinions**: Número de opiniones dejadas por los huéspedes a través de Tripadvisor
- **Number opinions excellent**: Número de opiniones excelentes dejadas por los huéspedes a través de Tripadvisor
- **Number opinions good**: Número de opiniones buenas dejadas por los huéspedes a través de Tripadvisor

- **Number opinions normal**: Número de opiniones normales dejadas por los huéspedes a través de Tripadvisor
- **Number opinions bad**: Número de opiniones malas dejadas por los huéspedes a través de Tripadvisor
- **Number opinions awful**: Número de opiniones pésimas dejadas por los huéspedes a través de Tripadvisor
- **Number QA**: Número de preguntas y respuestas de los usuarios
- **Nearby restaurants**: Número de restaurantes cercanos al hotel
- **Nearby attractions**: Número de atracciones turísticas cercanas al hotel
- **Zone**: Zona en la que se sitúa el hotel
- **Latitude/Longitude**: Latitud y longitud del hotel
- **Swimming pool**: Nos dice si el hotel dispone de este servicio
- **Bar**: Nos dice si el hotel dispone de este servicio
- **Restaurant**: Nos dice si el hotel dispone de este servicio
- **Breakfast**: Nos dice si el hotel dispone de este servicio
- **Gym**: Nos dice si el hotel dispone de este servicio
- **Reception 24h**: Nos dice si el hotel dispone de este servicio
- **Admit pets**: Nos dice si el hotel dispone de este servicio
- **Air conditioning**: Nos dice si el hotel dispone de este servicio
- **Strong box**: Nos dice si el hotel dispone de este servicio
- **Rooms**: Número de habitaciones del hotel
- **Suites**: Nos dice si el hotel dispone de este tipo de habitaciones
- **Sea View Rooms**: Nos dice si el hotel dispone de este tipo de habitaciones
- **Non-smoking Rooms**: Nos dice si el hotel dispone de este tipo de habitaciones
- **Landmark View Rooms**: Nos dice si el hotel dispone de este tipo de habitaciones
- **City View Rooms**: Nos dice si el hotel dispone de este tipo de habitaciones
- **Family Rooms**: Nos dice si el hotel dispone de este tipo de habitaciones
- **Style**: Estilo del hotel
- **Tripadvisor Clasification**: Calificación de Tripadvisor sobre la facilidad de realizar actividades y encontrar restaurantes a corta distancia del hotel.
- **Language Spanish**: Indica si el hotel habla este idioma o no
- **Language Catalan**: Indica si el hotel habla este idioma o no
- **Language French**: Indica si el hotel habla este idioma o no
- **Language English**: Indica si el hotel habla este idioma o no
- **Language Italian**: Indica si el hotel habla este idioma o no
- **Language Bulgarian**: Indica si el hotel habla este idioma o no
- **Language Russian**: Indica si el hotel habla este idioma o no
- **Language Portuguese**: Indica si el hotel habla este idioma o no
- **Prat Distance**: Distancia en kilómetros al aeropuerto del Prat
- **Timestamp**: Fecha/hora de recogida de los datos

Con la elección de este dataset, se pretende dar respuesta a diferentes preguntas sobre los hoteles, con el objetivo de ver si algunos atributos varían o no respecto a los barrios donde se sitúan los hoteles o entre diferentes rangos de precio. De esta forma, un posible turista puede ayudarse por este estudio para elegir entre los hoteles que más le convengan. Algunas de estas preguntas podrían ser las siguientes.

- ¿Hay relación entre las estrellas de un hotel y la valoración de los usuarios?
- ¿Hay diferencias en los precios de hoteles entre diferentes barrios de Barcelona?
- ¿Es cierto que los hoteles más caros ofrecen una mejor limpieza de sus instalaciones? ¿Y un mejor servicio?
- ¿Son más caros aquellos hoteles que admiten mascotas de los que no lo hacen? ¿Y los que tienen gimnasio?
- ¿Se hablan más idiomas en los hoteles de un barrio respecto a los demás? ¿Y en los más caros respecto a los demás?
- ¿La relación calidad precio es mayor en hoteles más baratos que en los caros?
- ¿Hay relación entre la puntuación que los usuarios atribuyen a la ubicación del hotel respecto al número

de restaurantes y atracciones cercanas?
- ¿Los hoteles con menos estrellas suelen tener más opiniones de usuarios que los de más estrellas?
- ¿Hay más habitaciones disponibles en algún barrio?

A parte, se intentará generar algún modelo de clasificación para predecir los rangos de precios de los hoteles según el resto de características.

2 Integración y selección de los datos de interés a analizar

El primer paso que debemos hacer es cargar el dataset que se va a utilizar. Lo leemos y utilizamos la función `str` para ver su estructura y una muestra de los datos.

```
library(stringr)
library(ggplot2)
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.0.3
```

```
df <- read.csv("tripadvisor_barcelona_hotels.csv", encoding = "UTF-8")
str(df, strict.width="wrap")
```

```
## 'data.frame':    1619 obs. of  50 variables:
## $ name : chr "Travelodge Barcelona Fira" "Four Points by Sheraton Barcelona
##    Diagonal" "Hotel Miramar Barcelona" "W Barcelona" ...
## $ stars : num 3 3 5 5 3 NA 4 3 4 4.5 ...
## $ score : num 3.5 4.5 4 4.5 4.5 4 5 4.5 4 5 ...
## $ score_location : num 4 4 4.5 4.5 4 4.5 5 4.5 4.5 5 ...
## $ score_cleaning : num 4.5 4.5 4.5 4.5 4.5 4 5 4.5 4.5 5 ...
## $ score_service : num 4 4.5 4 4.5 4.5 4 4.5 4.5 4 5 ...
## $ score_value_money : num 4 4.5 4 4 4.5 4.5 4.5 4.5 4 4.5 ...
## $ ranking : num 354 150 198 243 134 80 10 92 313 6 ...
## $ price : num NA 76 184 323 143 NA NA 98 126 160 ...
## $ price_range : chr "52 \200 - 166 \200" "79 \200 - 195 \200" "152 \200 -
##    283 \200" "287 \200 - 813 \200" ...
## $ opinions : num 234 1574 1298 7751 1404 ...
## $ opinions_excellent : num 77 685 708 4598 693 ...
## $ opinions_good : num 72 692 339 1789 566 ...
## $ opinions_normal : num 49 137 107 671 103 29 17 127 200 13 ...
## $ opinions_bad : num 19 31 66 329 31 11 4 21 56 4 ...
## $ opinions_awful : num 17 30 78 364 11 24 2 15 51 3 ...
## $ num_qa : num 15 99 28 182 71 10 23 44 34 94 ...
## $ nearby_restaurants : num 66 170 55 18 468 388 454 566 668 594 ...
## $ nearby_attractions : num 7 11 20 2 48 50 113 111 275 296 ...
## $ zone : chr "Barcelona" "Poblenou / Sant Martí" "Montjuic / Sants-Montjuïc"
##    "Ciutat Vella / El Port Vell / Barceloneta" ...
## $ latitude : num 41.4 41.4 41.4 41.4 41.4 ...
## $ longitude : num 2.13 2.2 2.17 2.19 2.15 ...
## $ has_swimming_pool : chr "False" "False" "True" "True" ...
## $ has_bar : chr "True" "True" "True" "True" ...
## $ has_restaurant : chr "True" "True" "True" "True" ...
## $ has_breakfast : chr "True" "True" "True" "True" ...
## $ has_gym : chr "False" "True" "True" "True" ...
## $ has_reception_24h : chr "True" "True" "True" "True" ...
## $ has_ac : chr "True" "True" "True" "True" ...
```

```
## $ has_strongbox : chr "False" "True" "True" "True" ...
## $ admits_pets : chr "False" "False" "False" "True" ...
## $ rooms : num 83 154 75 473 53 16 19 64 81 101 ...
## $ suites : chr "False" "False" "False" "True" ...
## $ sea_views_rooms : chr "False" "False" "True" "True" ...
## $ non_smoking_rooms : chr "True" "True" "True" "True" ...
## $ landmarks_views_rooms : chr "False" "False" "False" "False" ...
## $ city_views_rooms : chr "False" "False" "False" "False" ...
## $ family_rooms : chr "True" "False" "False" "False" ...
## $ style : chr "Familiar" "Ecológico" "Vistas al parque" "Vistas a la bahía" ...
## $ tripadvisor_clasification: chr "51 de 100" "94 de 100" "97 de 100" "75 de
##    100" ...
## $ language_spanish : chr "True" "True" "False" "True" ...
## $ language_catalan : chr "True" "True" "True" "False" ...
## $ language_french : chr "True" "True" "True" "False" ...
## $ language_english : chr "True" "True" "True" "False" ...
## $ language_italian : chr "True" "True" "False" "False" ...
## $ language_bulgarian : chr "True" "False" "False" "False" ...
## $ language_russian : chr "False" "False" "False" "False" ...
## $ language_portuguese : chr "False" "True" "False" "False" ...
## $ prat_distance : int 7 15 11 12 12 11 13 12 12 13 ...
## $ timestamp : chr "2020-11-02 15:37:45.938143" "2020-11-02 15:37:58.762559"
##    "2020-11-02 15:38:01.677586" "2020-11-02 15:38:06.467551" ...
```

En cuanto a la selección de datos, vamos a quedarnos con aquellos que tengan un rango de precio informado, ya que como vamos a explicar será una de las variable objetivo del análisis. En cuanto a los diferentes atributos, vamos a empezar eliminando las columnas `name` y `timestamp`, ya que la primera contiene el nombre del hotel, el cual no es necesario para los análisis y el segundo la fecha de la captura de los datos, que tampoco nos aporta información relativa al hotel.

Observamos también dos columnas relativas al precio. `price` contiene el precio por noche en el momento de la extracción del dato proporcionado por terceros (Booking, Expedia...). `price_range` es una estimación más general del precio del hotel y es proporcionado por la propia Tripadvisor. Además, la variable `price`, a parte de tener un carácter temporal, tiene muchos valores nulos, como veremos más tarde. Por este motivo, eliminaremos la variable `price` y obtendremos los precios directamente de la variable `price_range`.

```
# eliminamos las 3 variables mencionadas
df <- df[, !(names(df) %in% c("name", "timestamp", "price"))]
# eliminamos las filas que tengan nulos en price_range (string vacío)
df <- df[df$price_range != "",]
```

3 Limpieza de los datos

3.1 Preprocesado

El primer paso será convertir el tipo de las variables. Vamos a empezar convirtiendo a tipo factor las variables categóricas.

```
catergorical_vars <- c("has_swimming_pool", "has_bar", "has_restaurant", "has_breakfast",
                       "has_gym", "has_reception_24h" , "has_ac", "has_strongbox",
                       "admits_pets", "suites", "sea_views_rooms",  "non_smoking_rooms",
                       "landmarks_views_rooms", "city_views_rooms", "family_rooms",
                       "style", "language_spanish", "language_catalan", "language_french",
                       "language_english", "language_italian", "language_bulgarian",
```

```
                        "language_russian", "language_portuguese")
for(v in catergorical_vars){
  df[[v]] <- factor(df[[v]])
}
```

A continuación, debemos estandarizar la variable `price_range`. Esta variable son rangos de precios medios. El problema es que estos rangos no están predefinidos de ninguna manera. Para solucionar este problema, vamos a transformar la variable para obtener el precio medio del intervalo y luego discretizar la variable en los valores "Barato", "Medio", "Caro" y "Lujo" según las siguientes reglas:

- **Barato:** precio $<$ 60 euros
- **Medio:** 60 $>=$ precio $<$ 150
- **Caro:** 150 $>=$ precio $<$ 300
- **Lujo:** precio $>=$ 300

A continuación, se muestra la variable antes y después de esta transformación.

```
head(df$price_range, 3)
price_range_clean <- str_replace_all(df$price_range, "€|[:space:]|\\.", "")
price_range_clean <- str_split(price_range_clean, "-")
new_price_range <- rep(NA, length(price_range_clean))
for(i in 1:length(price_range_clean)){
  range <- as.list(price_range_clean[i][[1]])
  if(range[[1]] != ""){
    m <- mean(c(strtoi(range[[1]]), strtoi(range[[2]])))
    new_price_range[i] <- m
  }
}
df$price_range <- new_price_range
head(df$price_range, 3)

# discretizar
df$price_range_disc[df$price_range < 60] <- "Barato"
df$price_range_disc[(df$price_range >= 60) & (df$price_range < 150)] <- "Medio"
df$price_range_disc[(df$price_range >= 150) & (df$price_range < 300)] <- "Caro"
df$price_range_disc[df$price_range >= 300] <- "Lujo"
df$price_range_disc <- factor(df$price_range_disc,
                              levels = c("Barato", "Medio", "Caro", "Lujo"))
df$price_range <- NULL
head(df$price_range_disc, 3)
```

```
## [1] "52 \200 - 166 \200"  "79 \200 - 195 \200"  "152 \200 - 283 \200"
## [1] 109.0 137.0 217.5
## [1] Medio Medio Caro
## Levels: Barato Medio Caro Lujo
```

Por último, vamos a realizar un procesado para la variable `tripadvisor_clasification`. Es una puntuación sobre 100 y está expresada como un string de la forma "88 de 100", por ejemplo. El objetivo es obtener la puntuación y convertirla a entero.

```
head(df$tripadvisor_clasification, 5)
new_class <- rep(NA, nrow(df))
for(i in 1:nrow(df)){
  cl <- df$tripadvisor_clasification[i]
  index <- str_locate(cl, " de 100")[1]
  new_class[i] <- as.integer(str_sub(cl, 1, index))
```

```
}
df$tripadvisor_clasification <- new_class
head(df$tripadvisor_clasification, 5)
```

```
## [1] "51 de 100"  "94 de 100"  "97 de 100"  "75 de 100"  "100 de 100"
## [1]  51  94  97  75 100
```

3.2 Elementos nulos

A continuación, vamos a inspeccionar el dataset en busca de valores nulos.

```
colSums(is.na(df))
```

```
##                     stars                     score            score_location
##                       154                       103                       211
##            score_cleaning           score_service         score_value_money
##                       208                       190                       226
##                   ranking                  opinions        opinions_excellent
##                       105                       102                       103
##             opinions_good           opinions_normal              opinions_bad
##                       103                       103                       103
##            opinions_awful                   num_qa        nearby_restaurants
##                       103                       102                         5
##        nearby_attractions                      zone                  latitude
##                         5                         0                         3
##                 longitude         has_swimming_pool                   has_bar
##                         3                         0                         0
##            has_restaurant             has_breakfast                  has_gym
##                         0                         0                         0
##         has_reception_24h                    has_ac             has_strongbox
##                         0                         0                         0
##               admits_pets                     rooms                    suites
##                         0                       230                         0
##           sea_views_rooms         non_smoking_rooms     landmarks_views_rooms
##                         0                         0                         0
##          city_views_rooms              family_rooms                     style
##                         0                         0                         0
## tripadvisor_clasification          language_spanish          language_catalan
##                         5                         0                         0
##           language_french          language_english          language_italian
##                         0                         0                         0
##        language_bulgarian          language_russian       language_portuguese
##                         0                         0                         0
##             prat_distance          price_range_disc
##                         3                         0
```

Dependiendo de la variable, vamos a tratar los nulos de forma diferente.

- Las que hacen referencia al número de opiniones las vamos a imputar con un 0, ya que si es nulo es que no se ha encontrado ninguna opinión.
- Para la distancia al aeropuerto del Prat, latitud y longitud vamos a imputar los nulos con la mediana, ya que al estar en Barcelona es muy probable que estos valores sean similares al resto.
- Para las atracciones y restaurantes cercanos vamos a imputar también con la mediana (solo 5 nulos).

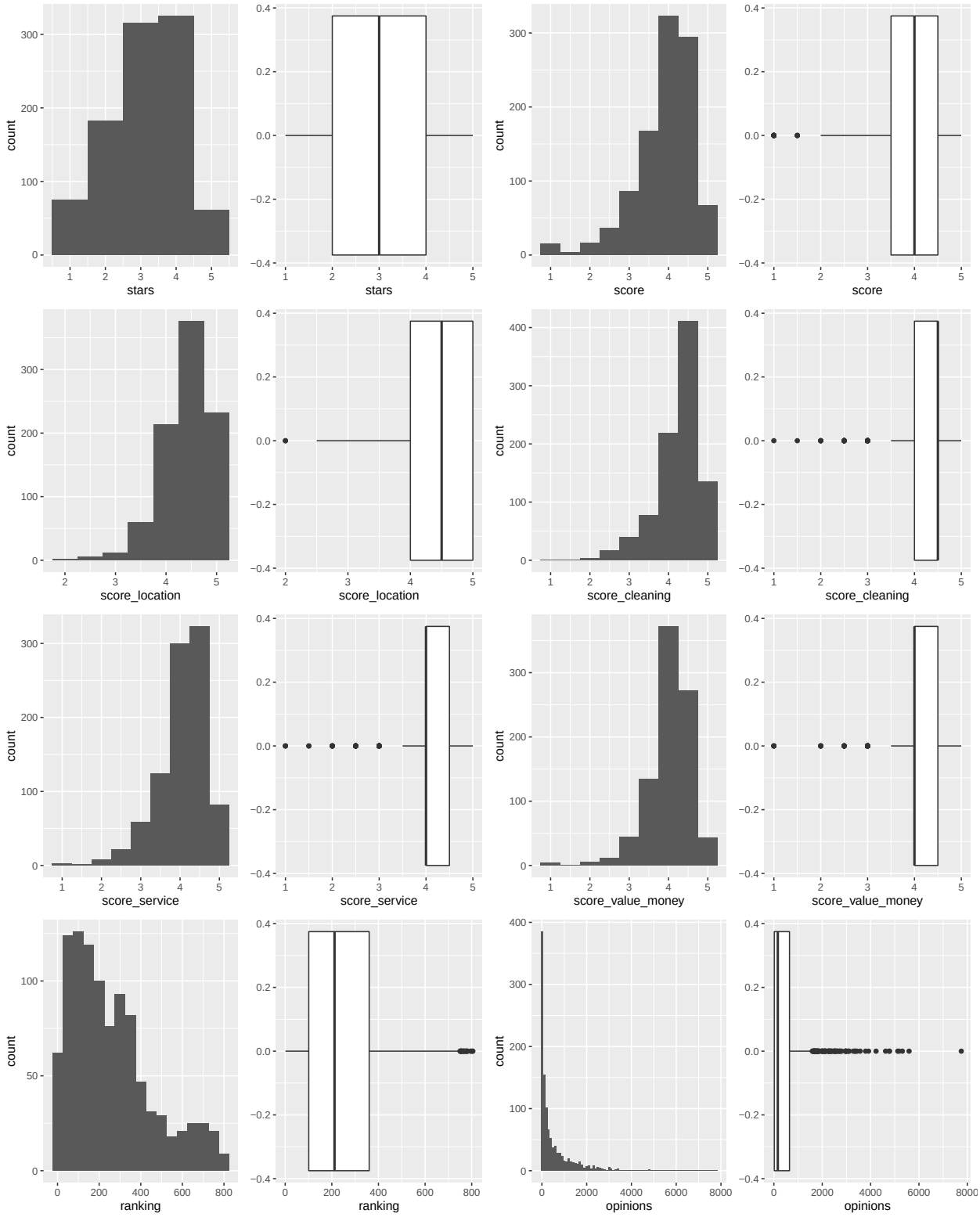
```r
# opinión
opinion_vars <- c("opinions", "opinions_excellent", "opinions_good", "opinions_normal",
                  "opinions_bad", "opinions_awful", "num_qa")
for(var in opinion_vars){
  df[[var]][is.na(df[[var]])] <- 0
}
# prat_distance, latitude, longitude, nearby_restaurants,  nearby_attractions
median_vars <- c("prat_distance", "latitude", "longitude", "nearby_restaurants",
                 "nearby_attractions")
for(var in median_vars){
  df[[var]][is.na(df[[var]])] <- median(df[[var]], na.rm = TRUE)
}
colSums(is.na(df))
```

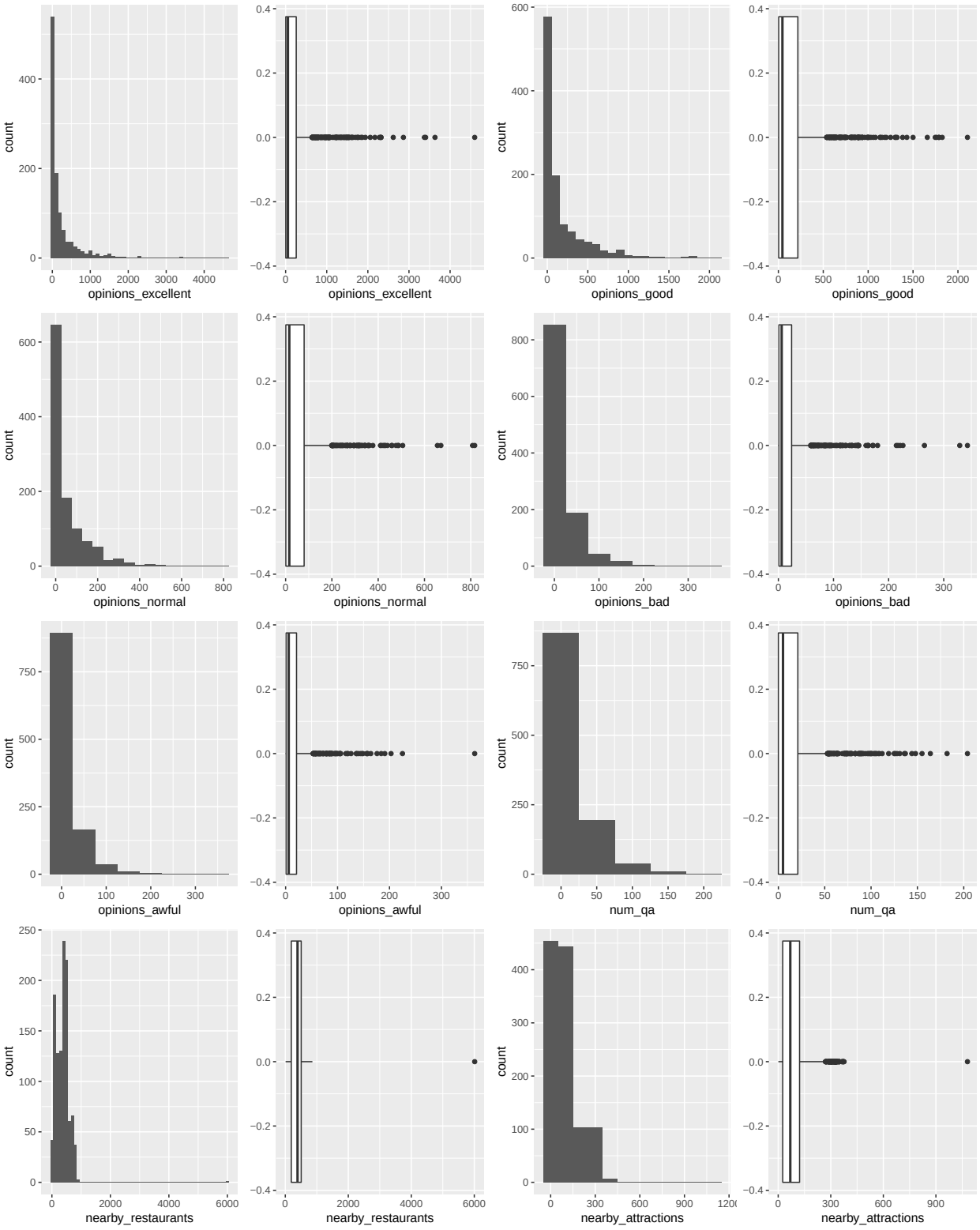
```
##                     stars                     score          score_location
##                       154                       103                       211
##            score_cleaning           score_service       score_value_money
##                       208                       190                       226
##                   ranking                  opinions      opinions_excellent
##                       105                         0                         0
##             opinions_good           opinions_normal             opinions_bad
##                         0                         0                         0
##            opinions_awful                    num_qa       nearby_restaurants
##                         0                         0                         0
##       nearby_attractions                      zone                  latitude
##                         0                         0                         0
##                 longitude         has_swimming_pool                   has_bar
##                         0                         0                         0
##            has_restaurant             has_breakfast                  has_gym
##                         0                         0                         0
##         has_reception_24h                    has_ac             has_strongbox
##                         0                         0                         0
##               admits_pets                     rooms                    suites
##                         0                       230                         0
##           sea_views_rooms         non_smoking_rooms     landmarks_views_rooms
##                         0                         0                         0
##          city_views_rooms              family_rooms                     style
##                         0                         0                         0
## tripadvisor_clasification          language_spanish          language_catalan
##                         5                         0                         0
##           language_french          language_english          language_italian
##                         0                         0                         0
##        language_bulgarian          language_russian       language_portuguese
##                         0                         0                         0
##             prat_distance          price_range_disc
##                         0                         0
```

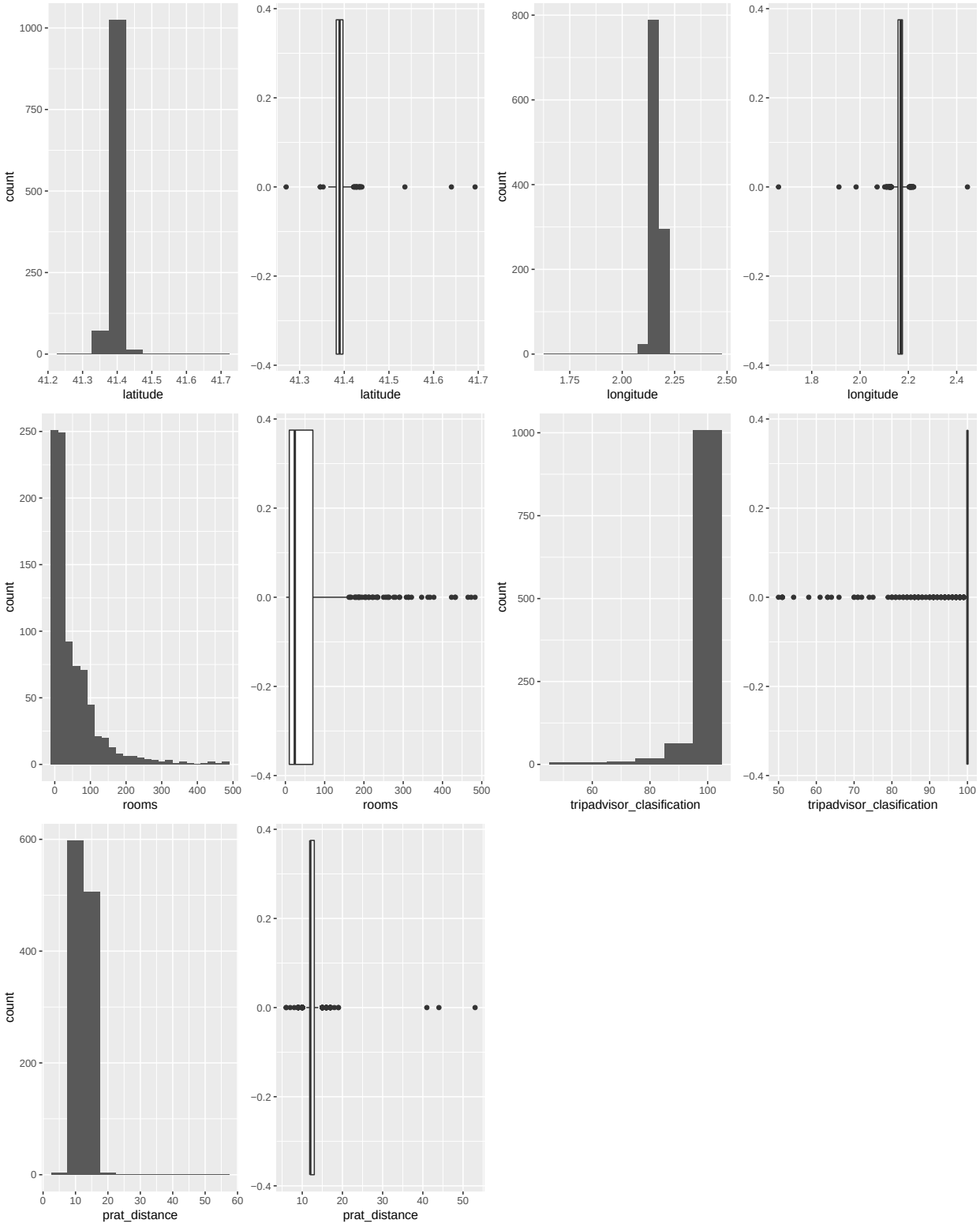
De momento no vamos a imputar el resto de nulos. Será en el momento que necesitemos usar estas columnas cuando eliminaremos aquellas filas que tengan NA's, ya que debido a que en algunas columnas el porcentaje de nulos está cerca del 20%, es mejor eliminarlas para no desvirtuar las posibles conclusiones. Sin embargo, no las vamos a eliminar ahora, ya que pueden tener información importante para el análisis del resto de variables.

3.3 Valores extremos

Vamos a inspeccionar las variables numéricas en busca de outliers. Nos ayudaremos de un histograma y de un boxplot de cada variable.




count
opinions_excellent
opinions_excellent
opinions_good
opinions_good
opinions_normal
opinions_normal
opinions_bad
opinions_bad
opinions_awful
opinions_awful
num_qa
num_qa
nearby_restaurants
nearby_restaurants
nearby_attractions
nearby_attractions


count
latitude
latitude
count
longitude
longitude
count
rooms
rooms
count
tripadvisor_clasification
tripadvisor_clasification
count
prat_distance
prat_distance

4 Análisis de los datos

5 Representación de los resultados a partir de tablas y gráficas

6 Conclusiones