Práctica 2: Booking Data Cleaning

Gerard Alcalde and Guillem Rochina

2022-12-24

## 1. DESCRIPCIÓN DEL DATASET

### ¿Por qué es importante y qué pregunta/problema pretende responder?

#### 1.1. Descripción general

El conjunto de datos ha tratar es el obtenido como resultado de la práctica anterior de Web Scraping. El dataset presenta los datos e indicadores más relevantes de cada uno de los hoteles encontrados en función de determinados criterios de búsqueda (ciudad, fecha de check-in, fecha de check-out, número de adultos, niños y habitaciones). En este dataset en concreto presentamos datos para distintas fechas (diciembre, marzo y junio), ciudades (barcelona, madrid, valencia), número de adultos, niños y habitaciones a fin de tener una muestra más amplia e informativa que la que nos daría una búsqueda con tan sólo unos parámetros fijos.

A continuación, se realiza una breve descripción de cada una de las variables disponibles antes de realizar la limpieza:

* **Name (str)**: Nombre del hotel.
* **City (str)**: Nombre de la ciudad donde se realiza la búsqueda.
* **Check-in (str)**: Fecha de entrada al hotel.
* **Check-out (str)**: Fecha de salida del hotel.
* **Adults (int)**: Nº de adultos para los que se realiza la reserva.
* **Num\_rooms (int)**: Nº de habitaciones reservadas.
* **Address (str)**: Dirección postal del hotel, nombre de calle, barrio de la ciudad, código postal, etc.
* **Hotel\_coordinates (str)**: Latitud y longitud de la ubicación del hotel.
* **Hotel\_score (int)**: Nota general que recibe el hotel por los usuarios.
* **Hotel\_scores (dict)**: Puntuaciones de cada una de las dimensiones que puede valorar un usuario.
* **Hotel description (list)**: Descripción aportada por el propietario del hotel
* **Features (list)**: Lista de servicios añadidos del hotel.
* **Room\_data (dict)**: Conjunto de diccionarios que recogen las características de cada una de las habitaciones disponibles en el hotel.
* **Page\_count (int)**: Posición en la que ha aparecido el hotel en el buscador.
* **Current\_page (int)**: Página en la que ha aparecido el hotel en el buscador.

#### 1.2. Importancia y problema a responder

Aparecer en los primeros resultados de una búsqueda de un usuario por internet se ha convertido en los últimos años en una de las preocupaciones principales de cualquier negocio que busca hacerse un hueco en el mundo digital. Este fenómeno explica la proliferación de herramientas como el SEO y el SEM, mediante las cuales se intenta innovar (más allá del pago de tarifas para aparecer en las primeras páginas de búsqueda) a fin de lograr un mejor “posicionamiento” en buscadores como Google o Bing. El negocio hotelero no es una excepción en este caso.

No obstante, el uso de dichas herramientas no resulta útil si nos enfrentamos a un buscador dentro de una página web, como el caso que nos ocupa con Booking.com. ¿Qué hacer entonces? ¿Cómo mejorar el posicionamiento dentro de la web?

En la web del metabuscador de alojamientos encontramos que define a dichas posiciones entre los resultados como “ranking”, descrito por ellos mismos como el orden en el que se muestran los alojamientos disponibles en los resultados de búsqueda. En su blog “Booking.com Partner Hub” indican que los resultados son ordenados según la relevancia en base a las preferencias particulares de cada “cliente”, incluyendo en este caso también las “dinámicas del mercado”, el “rendimiento del alojamiento”, etc. Variable que poca información revela al interesado en que sus hoteles/alojamientos aparezcan en posiciones más “atractivas” para este.

El objetivo de este dataset es obtener las características que hacen que un hotel esté mejor posicionado que otro. El posicionamiento de los hoteles suele ser algo estático y por ello con una sola lectura de la página web puede ser suficiente.

En concreto, en el mismo post analizado se destaca la relevancia de revisar regularmente las “condiciones flexibles”, los comentarios, los precios externos, las promociones, la puntuación de la página del alojamiento… a fin de lograr escalar puestos en la clasificación.

Dado el impacto que puede tener en el negocio de un hotel el aparecer en posiciones más inmediatas del buscador, el presente proyecto pretende ser la continuación de la práctica 1, ofreciendo un análisis inferencial que destaque que factores tienen un mayor peso en el posicionamiento de un alojamiento en Booking.com.

## 2. LIMPIEZA DE LOS DATOS

A fin de iniciar el proceso de limpieza de datos, en primer lugar se debe realizar la lectura del fichero csv que obtuvimos en la práctica anterior. Como resultado de la función de R-base “read.csv” obtenemos un objeto data.frame, que manipularemos en la presente sección.

# Lectura y breve descripción de los datos  
  
booking <- read.csv("hotels\_data.csv", header = TRUE)  
summary(booking)

## name city check.in adults   
## Length:3912 Length:3912 Length:3912 Min. :2.000   
## Class :character Class :character Class :character 1st Qu.:2.000   
## Mode :character Mode :character Mode :character Median :4.000   
## Mean :3.329   
## 3rd Qu.:5.000   
## Max. :5.000   
##   
## children check.out num\_rooms address   
## Min. :1.000 Length:3912 Min. :2.000 Length:3912   
## 1st Qu.:1.000 Class :character 1st Qu.:2.000 Class :character   
## Median :3.000 Mode :character Median :2.000 Mode :character   
## Mean :2.475 Mean :2.268   
## 3rd Qu.:3.000 3rd Qu.:3.000   
## Max. :3.000 Max. :3.000   
##   
## hotel\_coordinates hotel\_score hotel\_scores hotel\_description   
## Length:3912 Min. :-1.000 Length:3912 Length:3912   
## Class :character 1st Qu.: 8.000 Class :character Class :character   
## Mode :character Median : 8.400 Mode :character Mode :character   
## Mean : 8.259   
## 3rd Qu.: 8.800   
## Max. :10.000   
## NA's :8   
## features room\_data page\_count current\_page   
## Length:3912 Length:3912 Min. : 0.00 Min. :1.000   
## Class :character Class :character 1st Qu.: 24.00 1st Qu.:1.000   
## Mode :character Mode :character Median : 53.00 Median :3.000   
## Mean : 57.19 Mean :2.741   
## 3rd Qu.: 88.00 3rd Qu.:4.000   
## Max. :133.00 Max. :5.000   
##   
## in\_page\_count search\_date   
## Min. : 0.00 Length:3912   
## 1st Qu.: 6.00 Class :character   
## Median :13.00 Mode :character   
## Mean :12.66   
## 3rd Qu.:19.00   
## Max. :27.00   
##

Como observamos tanto con la función “summary” como con la función “class” nos muestran que a cada variable se le ha asignado o bien la clase de entero o la de string. En concreto, muchas de las variables interpretadas como string se tratan de diccionarios y listas, factor que deberemos tener en cuenta a la hora de trabajar con estas variables.

# Tipo de dato asignado a cada variable  
  
sapply(booking, function(x) class(x))

## name city check.in adults   
## "character" "character" "character" "integer"   
## children check.out num\_rooms address   
## "integer" "character" "integer" "character"   
## hotel\_coordinates hotel\_score hotel\_scores hotel\_description   
## "character" "numeric" "character" "character"   
## features room\_data page\_count current\_page   
## "character" "character" "integer" "integer"   
## in\_page\_count search\_date   
## "integer" "character"

#### 2.1. Transformación de las variables en el formato adecuado

Como se ha comentado, existen algunas variables que presentan todavía una estructura no válida para tratarlos estadísticamente o introducirlos en un modelo de data mining o machine learning. En esta sección nos encargamos de tratar dichas columnas a fin de obtener nuevas variables con un formato adecuado, eliminando a su vez cualquier informaión que no nos sea de utilidad.

En primer lugar, cabe destacar que no todas las columnas que tenemos disponibles en el dataset nos ofrecen información relevante para nuestro problema. Por tanto, a fin de reducir la dimensionalidad de este, nos encargamos de eliminar las columnas con información redundante.

# Seleccionamos con la función "select" de dplyr todas las columnas excepto name y search\_date, que nos van a ser de poca utilidad.  
  
booking <- booking %>%  
 select(-c("name", "search\_date"))

La columna hotel\_coordinates incluye tanto los valores de latitud como los de longitud, separados por una coma. Dado que ambos dos valores nos indica información distinta, pues la latitud nos proporcionar información de la posición en dirección norte o sur del ecuador y la longitud información de la posición en dirección este u oeste, en este caso hemos optado por separarlos en dos columnas distintas.

# Separamos los valores de la variable hotel coordinates en dos nuevas columnas, latitud y longitud  
# La columna hotel\_coordinates se elimina en el proceso  
  
booking <- booking %>%  
 separate(hotel\_coordinates, c("latitude", "longitude"), sep = ",")

## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 16 rows [202,  
## 203, 223, 224, 225, 226, 227, 664, 665, 765, 1598, 2455, 2672, 2789, 2993,  
## 2999].

Por su parte, “hotel\_scores” alberga diccionarios en cada uno de los registros. Cada diccionario consta del nombre de cada dimensión (caracterísitca a valorar) como clave y la valoración que recibe como valor. Dado que R lo ha interpretado como un string, nuestro enfoque en su limpieza se ha basado en separar la información con comas y, una vez obtenidas todas las columnas resultantes, extraer los datos numéricos de cada registro para quedarnos solo con los scores de cada dimensión.

# Definimos un vector con los nombres de las columnas a crear  
  
columns <- c("staff\_score", "facilities\_score", "cleanliness\_score", "comfort\_score", "value\_for\_money\_score", "location\_score", "free\_wifi\_score")  
  
# Separamos los strings (diccionarios) en función de las comas y creamos las columnas correspondientes. La primera columna creada es una NA porque todos los primeros registros en cada diccionario están vacíos (por un fallo en la extracción de datos). Estableciendo la columna como NA, indicamos a la función separate que ignore dichos datos y no cree columna alguna para ellos.  
  
booking <- booking %>%  
 separate(hotel\_scores, c(NA, "staff\_score", "facilities\_score", "cleanliness\_score", "comfort\_score", "value\_for\_money\_score", "location\_score", "free\_wifi\_score"), sep = ",")

## Warning: Expected 8 pieces. Missing pieces filled with `NA` in 423 rows [9, 80,  
## 105, 125, 141, 202, 203, 218, 223, 224, 225, 226, 227, 247, 303, 311, 322, 327,  
## 336, 370, ...].

# Extraemos los valores numéricos de cada columna con la función del paquete readr, "parse\_number"  
  
booking[columns] <- apply(booking[columns], 2, readr::parse\_number)

Dado que de las fechas tan sólo nos interesa el mes en el que se realiza la reserva (un alojamiento con unas características determinadas pdría estar mejor posicionado en un determinado mes que en otro) procedemos a extraer dicha información de las variables “check-in” y “check-out”. No obstante, por el momento no eliminamos dichas columnas y las transformamos a formato date por si a caso podemos hacer uso de ellas más adelante.

booking <- booking %>%  
 # Separamos el nombre del mes del resto de elementos del registro para check\_in y check\_out  
 separate(check.in, c(NA, "month\_in", NA), sep = "-", remove = FALSE) %>%  
 separate(check.out, c(NA, "month\_out", NA), sep = "-", remove = FALSE) %>%  
 # Modificamos la columna check\_in para que tenga el formato de fecha normal  
 separate(check.in, c("day", "month", "year"), sep = "-") %>%  
 mutate(month = ifelse(month == "March", "03", ifelse(month == "June", "06", "12"))) %>%  
 unite("check\_in", c(day, month, year), sep = "-") %>%  
 # Modificamos la columna check\_out para que tenga el formato de fecha normal  
 separate(check.out, c("day", "month", "year"), sep = "-") %>%  
 mutate(month = ifelse(month == "March", "03", ifelse(month == "June", "06", "12"))) %>%  
 unite("check\_out", c(day, month, year), sep = "-")  
  
# Transformamos las columnas check\_in y check\_out al format date  
booking["check\_in"] <- as.Date.character(booking$check\_in, "%d-%m-%Y")  
booking["check\_out"] <- as.Date(booking$check\_out, "%d-%m-%Y")

De la columna address tan sólo nos interesa el código postal (el barrio no se incluye en todos los registros a diferencia del CP, así que hemos optado por quedarnos con estos valores), por lo que extraemos dicha información para crear una nueva columna llamada “postal\_code”, eliminando la información redundante en el proceso.

# Hemos tenido que cambiar esta función para que el extract no diera problemas  
Sys.setlocale('LC\_ALL', 'C')

## [1] "C"

# Extraemos los valores de codigo postal de la columna addres.  
booking <- booking %>%  
 extract(address, c("postal\_code"), regex = "( [0-9]\* )")

Los features, entendidos como servicios adicionales más allá de la propia habitación, fueron guardados como una lista. No obstante, de nuevo nos encontramos como R lo ha identificado como string. Es por ello por lo que realizamos un tratamiento similar al de la columna de scores. En este caso creamos tantas variables dummy (columnas dicotómicas con valores de 1, en caso de que el feature exista en dicho registro, o 0, en caso de que no existe) como servicios adicionales de hotel consideramos interesantes.

# Definimos las features que consideramos que son interesantes. Utilizamos este formato para filtrar por expresión regular.  
feature\_list <- c("('Free Wifi')", "('Air conditioning')", "('24-hour front desk')", "('Safe')", "('Heating')", "('Elevator')", "('Private Bathroom')", "('Non-smoking rooms')", "('Aparments')", "('City view')","('Kitchen')", "('Pet Friendly')", "('Swimming pool')", "('Balcony')")  
  
# Creamos un bucle en el que se recorre cada elemento de la lista de features  
for (feature in feature\_list){  
 # Se eliminan los parentésis y comillas de la variable local feature para crear el nombre de la columna.  
 col\_name <- str\_replace(str\_replace(str\_to\_lower(str\_extract(feature, "([A-Z][a-z]\*( |-)?[A-Z]?[a-z]\* ? ?[a-z]\*)")), " ", "\_"), "-", "\_")  
 booking <- booking %>%  
 # Extraemos el nombre del feature de cada uno de los registros (strings). En caso de que no encuentre ningún valor devuelve un NA.  
 extract(features, c(col\_name), regex = feature, remove = FALSE) %>%  
 # Transformamos la columna que acaba de ser creada para que indique con un 1 si el registro tenia dicho servicio y 0 si el valor era NA (no tenida dicho servicio)  
 mutate\_(.dots = setNames(list(paste0("as.integer(!is.na(",col\_name,"))")), col\_name))  
}

## Warning: `mutate\_()` was deprecated in dplyr 0.7.0.  
## i Please use `mutate()` instead.  
## i See vignette('programming') for more help  
## i The deprecated feature was likely used in the dplyr package.  
## Please report the issue at <]8;;https://github.com/tidyverse/dplyr/issueshttps://github.com/tidyverse/dplyr/issues]8;;>.

# Eliminamos la columna de features para finalizar  
booking <- booking %>%  
 select(-c("features"))

#### 2.2. Ceros y elementos vacíos

booking <- booking[booking$latitude != "",]

booking <- booking[booking$latitude != "",]

booking %>%  
 distinct() %>%  
 group\_by(city, month\_in) %>%  
 count()

## # A tibble: 9 x 3  
## # Groups: city, month\_in [9]  
## city month\_in n  
## <chr> <chr> <int>  
## 1 Barcelona December 654  
## 2 Barcelona June 297  
## 3 Barcelona March 398  
## 4 Madrid December 568  
## 5 Madrid June 380  
## 6 Madrid March 296  
## 7 Valencia December 522  
## 8 Valencia June 297  
## 9 Valencia March 323

#### 2.3. Valores extremos

#### 2.4. Exportación de los datos postprocesados

## 3. INTEGRACIÓN Y SELECCIÓN

En primer lugar, hemos obtado por añadir datos de vuelos por varias razones:

* Un mayor número de vuelos en un mes podría aumentar las búsquedas de hoteles cerca del aeropuerto.
* Un mayor número de vuelos en un mes indica más turismo internacional (en lugar de sólo turismo nacional). Un tipo de turismo que podría estar buscando hoteles con características diferentes a los turistas nacionales.
* Un mayor número de vuelos en un mes indica más turismo general, lo que puede influir en las estrategias de promoción y ofertas de determinados hoteles, influyendo en última instancia en su posición en el buscador Booking.

flights <- read.csv("avia\_tf\_apal\_linear.csv.gz")  
summary(flights)

## DATAFLOW LAST.UPDATE freq unit   
## Length:1502955 Length:1502955 Length:1502955 Length:1502955   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## tra\_meas airline rep\_airp TIME\_PERIOD   
## Length:1502955 Length:1502955 Length:1502955 Length:1502955   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## OBS\_VALUE OBS\_FLAG   
## Min. : 0 Mode:logical   
## 1st Qu.: 0 NA's:1502955   
## Median : 264   
## Mean : 75333   
## 3rd Qu.: 3166   
## Max. :289204357

proc\_flights <- flights %>%   
 # Código OACI/ICAO: Barcelona --> ES\_LEBL; Valencia --> ES\_LEVC; Madrid: ES\_LEMD  
 filter(rep\_airp %in% c("ES\_LEBL", "ES\_LEVC", "ES\_LEMD")) %>%   
 filter(tra\_meas == "PAS\_CRD") %>%  
 filter(freq == "M") %>%  
 separate(TIME\_PERIOD, c("YEAR", "MONTH"), sep = "-") %>%  
 filter(YEAR %in% c("2022", "2021", "2020", "2019", "2018") & MONTH %in% c("03", "06", "12")) %>%  
 filter(airline == "TOTAL") %>%  
 mutate(month\_name = ifelse(MONTH == "03", "March", ifelse(MONTH == "06", "June", "December")),  
 city\_airp = ifelse(rep\_airp == "ES\_LEBL", "Barcelona", ifelse(rep\_airp == "ES\_LEVC", "Valencia", "Madrid"))) %>%  
 group\_by(rep\_airp, month\_name) %>%  
 mutate(mean\_fligths = mean(OBS\_VALUE)) %>%  
 ungroup() %>%  
 select(c(city\_airp, month\_name, mean\_fligths))  
  
final\_flights <- unique(proc\_flights)

booking <- booking %>%  
 merge(final\_flights, by.x = c("city", "month\_in"), by.y = c("city\_airp", "month\_name"))

## ANÁLISIS DE LOS DATOS

## RESOLUCIÓN DEL PROBLEMA