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? Asimismo, la fijación de precios o “pricing” puede ser una tardea ardua en los últimos años, pues la aparición de apartamentos vacacionales a través de plataformas como AirBnB ha incrementado de manera significativa la competencia a la que se deben enfrentar los propietarios de los hoteles, lo que influye en última instancia los precios a establecer por sus servicios. Dicho esto, ¿Qué precios se deben fijar dadas las características de un hotel? ¿Añadir un servicio adicional permitiría a un hotel elevar los precios de sus habitaciones?. Ante las preguntas presentadas en los párrafos anteriores, el objetivo de este proyecto es dual. Por un lado se busca obtener las características que hacen que un hotel esté mejor posicionado que otro en el metabuscador de Booking.com y, por el otro, se persigue desarrollar un modelo que indique qué precios se deberían establecer dadas las características de una habitación y qué elementos pueden incrementar el precio de forma relevante. En conclusión, el presente proyecto pretende dar continuidad a la Práctica 1, explotando los datos que se obtuvieron mediante web scrapping para ayudar a superar dos de los desafíos más relevantes para el sector hotelero.

## 2. LIMPIEZA DE LOS DATOS

Dado que muchas de las variables que desecharemos en la sección de selección e integración las vamos a utilizar en a continuación, procedemos a realizar previamente la limpieza de los datos. A fin de iniciar el proceso, 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("../data/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   
##

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 la limpieza de estas variables.

# Tipo de dato asignado a cada variable  
  
res <- sapply(booking, function(x) class(x))  
kable(data.frame(var=names(res), clase=as.vector(res)))

| var | clase |
| --- | --- |
| name | character |
| city | character |
| check.in | character |
| adults | integer |
| children | integer |
| check.out | character |
| num\_rooms | integer |
| address | character |
| hotel\_coordinates | character |
| hotel\_score | numeric |
| hotel\_scores | character |
| hotel\_description | character |
| features | character |
| room\_data | character |
| page\_count | integer |
| current\_page | integer |
| in\_page\_count | integer |
| search\_date | 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. En esta sección nos encargamos de tratar dichas columnas a fin de obtener nuevas variables con un formato adecuado, eliminando a su vez cualquier información que no nos sea de utilidad.

La columna hotel\_coordinates incluye tanto los valores de latitud como los de longitud, separados por una coma. Dado que ambos dos valores nos indican 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 = ",", remove = FALSE)

## 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].

booking$latitude <- as.numeric(booking$latitude)  
booking$longitude <- as.numeric(booking$longitude)

Por su parte, hotel\_scores alberga un diccionario en cada uno de los registros. Cada diccionario consta del nombre de cada dimensión (característica 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 en base a las comas como separador y, una vez obtenidas todas las columnas resultantes, extraer los datos numéricos de cada registro para quedarnos solo con los scores de cada dimensión con la función “parse\_numbers”.

# 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 = ",", remove = FALSE)

## 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 podría estar mejor posicionado en un determinado mes que en otro o tener un precio distinto) procedemos a extraer dicha información de las variables check-in y check-out. De hecho, dado que los datos que hemos extraído corresponden al mismo mes en todos los registros, tan sólo nos quedaremos con una columna de mes. Asimismo, no crearemos una columna is\_june, pues la dejaremos como caso base, es decir, cuando no sea ninguna de las otras dos opciones (o Diciembre o Marzo).

booking <- booking %>%  
 # Separamos el nombre del mes del resto de elementos del registro para check\_in y check\_out. Eliminamos las columnas en el proceso.  
 separate(check.in, c(NA, "month", NA), sep = "-", remove = FALSE) %>%  
 mutate(is\_december = ifelse(month == "December", 1, 0),  
 is\_march = ifelse(month == "March", 1, 0))

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.

# 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 utilizando expressiones regulares  
booking <- booking %>%  
 extract(address, c("postal\_code"), regex = "( [0-9]{5} )", remove = FALSE)

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 exista) 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]?[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;;>.

Por otro lado, nos encontramos con la columna que incluye la descripción del hotel, esta descripción es muy amplia y un análisis profundo de ella requeriría de técnicas de NLP, las cuales no son el objetivo de esta práctica. No obstante, consideramos que la longitud de la descripción del hotel sí que puede tener una relación con la posición en el buscador, por lo que transformaremos esta variable a una nueva variable que contenga el número total de palabras que contiene la descripción.

# Evaluamos la longitud de la descripción del hotel  
booking$length\_description <- lengths(gregexpr("\\W+", booking$hotel\_description)) + 1

Finalmente, en la columna room\_data, se incluye mucha información referente a los tipos de habitaciones disponibles, precios, características, etc. Esta se había extraído de esta forma con el objetivo de procesarla mediante un diccionario de Python. No obstante, al haber optado por un procesado con R este se complicará un poco más. En esta columna tenemos distintos datos que pueden ser muy relevantes, de los cuales extraeremos los siguientes: • min\_price: Precio de la habitación más económica. • max\_price: Precio de las habitaciones más caras del hotel. • suite\_available: Una de las habitaciones es de tipo suite. • is\_apartment: Una de las habitaciones es de tipo apartamento. • free\_cancelation\_available: Alguna de las habitaciones tiene cancelación gratuita.

# Función que encuentra el precio mínimo de la habitación dado un string con la información de las habitaciones  
find\_min\_price <- function(text) {  
 if (text=="{}"){  
 return(NA)  
 }  
 a <- gregexpr("([0-9]\*', 'room\_capacity)", text)  
   
 min\_price <- NA  
 for (value in a[[1]]){  
 substring <- substr(text, value, value+20)  
 price <- as.numeric(strsplit(substring, "'")[[1]][1])  
 if (is.na(min\_price)) {  
 min\_price <- price  
 }  
   
 if (price < min\_price) {  
 min\_price <- price  
 }   
 }  
 if (min\_price == 0){  
 return(NA)  
 }  
 else {  
 return(min\_price)  
 }  
}  
  
# Aplicamos la función a todos los valores  
min\_price\_vector <- c()  
for (i in seq(1, length(booking$room\_data))) {  
 min\_price\_vector <- append(min\_price\_vector, find\_min\_price(booking$room\_data[i]))  
}  
  
# Guardamos los resultados en una nueva columna.  
booking$min\_price <- min\_price\_vector  
  
# Función que encuentra el precio máximo de la habitación dado un string con la información de las habitaciones  
find\_max\_price <- function(text) {  
 if (text=="{}"){  
 return(NA)  
 }  
 a <- gregexpr("([0-9]\*', 'room\_capacity)", text)  
   
 max\_price <- NA  
 for (value in a[[1]]){  
 substring <- substr(text, value, value+20)  
 price <- as.numeric(strsplit(substring, "'")[[1]][1])  
 if (is.na(max\_price)) {  
 max\_price <- price  
 }  
   
 if (price > max\_price) {  
 max\_price <- price  
 }   
 }  
 if (max\_price == 0){  
 return(NA)  
 }  
 else {  
 return(max\_price)  
 }  
}  
  
# Aplicamos la función a todos los valores  
max\_price\_vector <- c()  
for (i in seq(1, length(booking$room\_data))) {  
 max\_price\_vector <- append(max\_price\_vector, find\_max\_price(booking$room\_data[i]))  
}  
  
# Guardamos los resultados en una nueva columna.  
booking$max\_price <- max\_price\_vector  
  
# Buscamos si hay una habitación en suite  
is\_suite\_vector <- c()  
for (i in seq(1, length(booking$room\_data))) {  
 is\_suite <- gregexpr("(suite)", booking$room\_data[i])[[1]][1]  
   
 if (is\_suite==-1) {  
 is\_suite\_vector <- append(is\_suite\_vector, 0)  
 } else {  
 is\_suite\_vector <- append(is\_suite\_vector, 1)  
 }  
}  
  
# Guardamos los resultados en una nueva columna.  
booking$is\_suite <- is\_suite\_vector  
  
# Buscamos si hay opción de apartamento  
is\_apartment\_vector <- c()  
for (i in seq(1, length(booking$room\_data))) {  
 is\_apartment <- gregexpr("(Apartment)", booking$room\_data[i])[[1]][1]  
   
 if (is\_apartment==-1) {  
 is\_apartment\_vector <- append(is\_apartment\_vector, 0)  
 } else {  
 is\_apartment\_vector <- append(is\_apartment\_vector, 1)  
 }  
}  
  
# Guardamos los resultados en una nueva columna.  
booking$is\_apartment <- is\_apartment\_vector  
  
# Buscamos si tiene cancelación gratuita  
free\_cancelation\_vector <- c()  
for (i in seq(1, length(booking$room\_data))) {  
 free\_cancelation <- gregexpr("(Free cancellation)", booking$room\_data[i])[[1]][1]  
   
 if (free\_cancelation==-1) {  
 free\_cancelation\_vector <- append(free\_cancelation\_vector, 0)  
 } else {  
 free\_cancelation\_vector <- append(free\_cancelation\_vector, 1)  
 }  
}  
  
# Guardamos los resultados en una nueva columna.  
booking$has\_free\_cancelation <- free\_cancelation\_vector

#### 2.2. Ceros y elementos vacíos

Dando paso a el tratamiento de ceros y elementos vacios, lo primero que debemos hacer es identificar aquellas variables en las que encontramos valores faltantes. Como observamos, “postal\_code” (19), “longitude” (16), “hotel\_score” (8), min (45) y max\_price (17) y el resto de variables relacionadas con los scores (60) presentan valores faltantes, destacando free\_wifi\_score con 423 registros faltantes, consecuencia con toda seguridad de la ausencia de este servicio en dichos hoteles.

# Comprobamos que columnas tienen datos faltantes  
kable(sapply(booking, function(y) sum(is.na(y))))

|  | x |
| --- | --- |
| name | 0 |
| city | 0 |
| check.in | 0 |
| month | 0 |
| adults | 0 |
| children | 0 |
| check.out | 0 |
| num\_rooms | 0 |
| address | 0 |
| postal\_code | 19 |
| hotel\_coordinates | 0 |
| latitude | 16 |
| longitude | 16 |
| hotel\_score | 8 |
| hotel\_scores | 0 |
| staff\_score | 60 |
| facilities\_score | 60 |
| cleanliness\_score | 60 |
| comfort\_score | 60 |
| value\_for\_money\_score | 60 |
| location\_score | 60 |
| free\_wifi\_score | 423 |
| hotel\_description | 0 |
| features | 0 |
| balcony | 0 |
| swimming\_pool | 0 |
| pet\_friendly | 0 |
| kitchen | 0 |
| city\_view | 0 |
| aparments | 0 |
| non\_smoking\_rooms | 0 |
| private\_bathroom | 0 |
| elevator | 0 |
| heating | 0 |
| safe | 0 |
| air\_conditioning | 0 |
| free\_wifi | 0 |
| room\_data | 0 |
| page\_count | 0 |
| current\_page | 0 |
| in\_page\_count | 0 |
| search\_date | 0 |
| is\_december | 0 |
| is\_march | 0 |
| as.integer(!is.na(NA)) | 0 |
| length\_description | 0 |
| min\_price | 45 |
| max\_price | 17 |
| is\_suite | 0 |
| is\_apartment | 0 |
| has\_free\_cancelation | 0 |

Por lo que respecta a la variable “postal\_code” encontramos que en todos aquellos registros donde se encuentan NAs, también hallamos datos faltantes en la mayoría del resto de columnas. Por tanto, dado que dichos registros no nos aportan información de valor, procedemos a eliminarlos del dataset.

# Comprobamos algunos de los registros donde hay datos faltantes de postal\_code  
kable(booking[is.na(booking$postal\_code),] %>%  
 dplyr::select(c("city", "month", "postal\_code", "longitude", "hotel\_score", "staff\_score")) %>%  
 head(3))

|  | city | month | postal\_code | longitude | hotel\_score | staff\_score |
| --- | --- | --- | --- | --- | --- | --- |
| 202 | Barcelona | December | NA | NA | -1 | NA |
| 203 | Barcelona | December | NA | NA | -1 | NA |
| 223 | Barcelona | December | NA | NA | -1 | NA |

# Filtramos los datos para eliminar aquellos registros que tienen NAs en postal\_code  
booking <- booking[!is.na(booking$postal\_code),]

Si comprobamos de nuevo qué columnas presentan todavía datos faltantes, observamos que en algunos casos (como “longitude”) hemos logrado eliminar todos los datos faltantes, mientras que en el resto se han reducido en 16 unidades el número de registros con NAs, pues como se había comentado todos aquellos registros con ausencia de postal\_code, también carecía del resto de datos.

# Comprobamos que columnas tienen datos faltantes  
kable(sapply(booking, function(y) sum(is.na(y))))

|  | x |
| --- | --- |
| name | 0 |
| city | 0 |
| check.in | 0 |
| month | 0 |
| adults | 0 |
| children | 0 |
| check.out | 0 |
| num\_rooms | 0 |
| address | 0 |
| postal\_code | 0 |
| hotel\_coordinates | 0 |
| latitude | 0 |
| longitude | 0 |
| hotel\_score | 5 |
| hotel\_scores | 0 |
| staff\_score | 44 |
| facilities\_score | 44 |
| cleanliness\_score | 44 |
| comfort\_score | 44 |
| value\_for\_money\_score | 44 |
| location\_score | 44 |
| free\_wifi\_score | 407 |
| hotel\_description | 0 |
| features | 0 |
| balcony | 0 |
| swimming\_pool | 0 |
| pet\_friendly | 0 |
| kitchen | 0 |
| city\_view | 0 |
| aparments | 0 |
| non\_smoking\_rooms | 0 |
| private\_bathroom | 0 |
| elevator | 0 |
| heating | 0 |
| safe | 0 |
| air\_conditioning | 0 |
| free\_wifi | 0 |
| room\_data | 0 |
| page\_count | 0 |
| current\_page | 0 |
| in\_page\_count | 0 |
| search\_date | 0 |
| is\_december | 0 |
| is\_march | 0 |
| as.integer(!is.na(NA)) | 0 |
| length\_description | 0 |
| min\_price | 29 |
| max\_price | 1 |
| is\_suite | 0 |
| is\_apartment | 0 |
| has\_free\_cancelation | 0 |

A continuación, analizamos los datos faltantes de hotel\_score. En este caso, encontramos que la ausencia de datos se registra de dos formas distintas, o bien con la introducción de NAs o bien con el valor -1. Por tanto, dado que el tratamiento en ambos casos va a ser el mismo, transformamos todos los valores -1 en NAs para poder llevar a cabo el siguiente proceso de inputación.

# Comprobamos algunos de los registros donde hay datos faltantes de hotel\_score  
booking[is.na(booking$hotel\_score),] %>%  
 dplyr::select(c("city", "month", "hotel\_score", "staff\_score", "location\_score", "longitude", "facilities\_score", "cleanliness\_score", "comfort\_score")) %>%  
 head(5)

## city month hotel\_score staff\_score location\_score longitude  
## 310 Barcelona December NA 9.4 9.4 2.1787906  
## 844 Valencia December NA 7.7 9.5 -0.3775288  
## 958 Valencia December NA 9.2 9.3 -0.3247243  
## 1210 Barcelona March NA 8.1 8.6 2.1429321  
## 2708 Barcelona June NA 8.1 8.6 2.1429321  
## facilities\_score cleanliness\_score comfort\_score  
## 310 8.6 8.9 8.9  
## 844 7.2 7.5 7.6  
## 958 9.1 9.3 9.3  
## 1210 7.3 7.7 7.7  
## 2708 7.3 7.7 7.7

# Comprobamos algunos de los registros donde hay datos faltantes de hotel\_score  
kable(booking[booking$hotel\_score == "-1",] %>%  
 dplyr::select(c("city", "month", "hotel\_score", "staff\_score", "location\_score", "longitude", "facilities\_score", "cleanliness\_score", "comfort\_score")) %>%  
 tail(5))

|  | city | month | hotel\_score | staff\_score | location\_score | longitude | facilities\_score | cleanliness\_score | comfort\_score |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 3690 | Madrid | June | -1 | NA | NA | -3.699663 | NA | NA | NA |
| 3736 | Madrid | March | -1 | NA | NA | -3.709937 | NA | NA | NA |
| 3855 | Madrid | March | -1 | NA | NA | -3.700559 | NA | NA | NA |
| 3877 | Madrid | March | -1 | NA | NA | -3.699663 | NA | NA | NA |
| 3909 | Madrid | March | -1 | NA | NA | -3.699676 | NA | NA | NA |

kable(booking[booking$hotel\_score == "-1",] %>%  
 dplyr::select(c("city", "month", "hotel\_score", "staff\_score", "free\_wifi\_score", "free\_wifi")) %>%  
 tail(5))

|  | city | month | hotel\_score | staff\_score | free\_wifi\_score | free\_wifi |
| --- | --- | --- | --- | --- | --- | --- |
| 3690 | Madrid | June | -1 | NA | NA | 1 |
| 3736 | Madrid | March | -1 | NA | NA | 0 |
| 3855 | Madrid | March | -1 | NA | NA | 1 |
| 3877 | Madrid | March | -1 | NA | NA | 1 |
| 3909 | Madrid | March | -1 | NA | NA | 1 |

# Tranformamos los valores -1 en NAs para poder llevar a cabo el proceso de inputación correctamente.  
booking$hotel\_score <- ifelse(booking$hotel\_score == "-1", NA, booking$hotel\_score)

Dado que en este caso tan sólo encontramos datos faltantes en las columnas de score, no consideramos eliminar estos registros, sino llevar a cabo una imputación. Aunque se podría haber utilizado una metología más simple como el uso de la media o la mediana, hemos optado por hacer uso de un método más complejo, basado en la similitud entre registros, es decir, una imputación mediante el algoritmo de K vecinos más próximos. En este caso, la función se encarga de devolver los mismos valores en caso de que ya existan en el dataset y el valor de los registros más “similares” en el caso de un registro con NAs.

# Imputamos los datos con K Neirest Neighbours  
booking$hotel\_score <- kNN(booking)$hotel\_score  
booking$staff\_score <- kNN(booking)$staff\_score   
booking$facilities\_score <- kNN(booking)$facilities\_score  
booking$cleanliness\_score <- kNN(booking)$cleanliness\_score  
booking$comfort\_score <- kNN(booking)$comfort\_score  
booking$value\_for\_money\_score <- kNN(booking)$value\_for\_money\_score  
booking$location\_score <- kNN(booking)$location\_score  
booking$free\_wifi\_score <- kNN(booking)$free\_wifi\_score

En el caso de los precios, encontramos NAs en aquellos registros que presentaban valores iguales a 0 (no son valores lógicos). En este caso también encontramos que los únicos valores faltantes en estos registros corresponde a precio, por lo que procedemos de la misma forma que con los scores.

kable(booking[is.na(booking$min\_price),] %>%  
 dplyr::select(c("city", "month", "hotel\_score", "max\_price", "min\_price")) %>%  
 tail(5))

|  | city | month | hotel\_score | max\_price | min\_price |
| --- | --- | --- | --- | --- | --- |
| 3620 | Madrid | June | 9.5 | 903 | NA |
| 3698 | Madrid | June | 8.2 | 964 | NA |
| 3846 | Madrid | March | 8.8 | 981 | NA |
| 3847 | Madrid | March | 8.6 | 960 | NA |
| 3849 | Madrid | March | 8.9 | 988 | NA |

# Imputamos los datos con K Neirest Neighbours  
booking$max\_price <- kNN(booking)$max\_price  
booking$min\_price <- kNN(booking)$min\_price

Una última comprobación nos revela que el dataset ya no cuenta con datos faltantes, con independencia del formato que pudieran presentar (strings vacías, valores de -1 en integers, o simplemente NAs).

# Comprobamos que columnas tienen datos faltantes (si los hay)  
kable(sapply(booking, function(y) sum(is.na(y))))

|  | x |
| --- | --- |
| name | 0 |
| city | 0 |
| check.in | 0 |
| month | 0 |
| adults | 0 |
| children | 0 |
| check.out | 0 |
| num\_rooms | 0 |
| address | 0 |
| postal\_code | 0 |
| hotel\_coordinates | 0 |
| latitude | 0 |
| longitude | 0 |
| hotel\_score | 0 |
| hotel\_scores | 0 |
| staff\_score | 0 |
| facilities\_score | 0 |
| cleanliness\_score | 0 |
| comfort\_score | 0 |
| value\_for\_money\_score | 0 |
| location\_score | 0 |
| free\_wifi\_score | 0 |
| hotel\_description | 0 |
| features | 0 |
| balcony | 0 |
| swimming\_pool | 0 |
| pet\_friendly | 0 |
| kitchen | 0 |
| city\_view | 0 |
| aparments | 0 |
| non\_smoking\_rooms | 0 |
| private\_bathroom | 0 |
| elevator | 0 |
| heating | 0 |
| safe | 0 |
| air\_conditioning | 0 |
| free\_wifi | 0 |
| room\_data | 0 |
| page\_count | 0 |
| current\_page | 0 |
| in\_page\_count | 0 |
| search\_date | 0 |
| is\_december | 0 |
| is\_march | 0 |
| as.integer(!is.na(NA)) | 0 |
| length\_description | 0 |
| min\_price | 0 |
| max\_price | 0 |
| is\_suite | 0 |
| is\_apartment | 0 |
| has\_free\_cancelation | 0 |

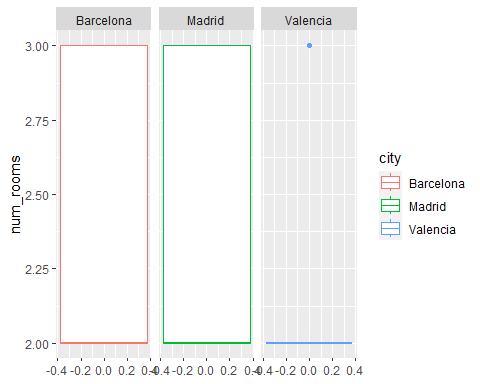
#Comprobamos que columnas tienen datos faltantes (si los hay)  
kable(sapply(booking, function(y) sum(y == "")))

|  | x |
| --- | --- |
| name | 0 |
| city | 0 |
| check.in | 0 |
| month | 0 |
| adults | 0 |
| children | 0 |
| check.out | 0 |
| num\_rooms | 0 |
| address | 0 |
| postal\_code | 0 |
| hotel\_coordinates | 0 |
| latitude | 0 |
| longitude | 0 |
| hotel\_score | 0 |
| hotel\_scores | 0 |
| staff\_score | 0 |
| facilities\_score | 0 |
| cleanliness\_score | 0 |
| comfort\_score | 0 |
| value\_for\_money\_score | 0 |
| location\_score | 0 |
| free\_wifi\_score | 0 |
| hotel\_description | 0 |
| features | 0 |
| balcony | 0 |
| swimming\_pool | 0 |
| pet\_friendly | 0 |
| kitchen | 0 |
| city\_view | 0 |
| aparments | 0 |
| non\_smoking\_rooms | 0 |
| private\_bathroom | 0 |
| elevator | 0 |
| heating | 0 |
| safe | 0 |
| air\_conditioning | 0 |
| free\_wifi | 0 |
| room\_data | 0 |
| page\_count | 0 |
| current\_page | 0 |
| in\_page\_count | 0 |
| search\_date | 0 |
| is\_december | 0 |
| is\_march | 0 |
| as.integer(!is.na(NA)) | 0 |
| length\_description | 0 |
| min\_price | 0 |
| max\_price | 0 |
| is\_suite | 0 |
| is\_apartment | 0 |
| has\_free\_cancelation | 0 |

#### 2.3. Valores extremos

Los outliers son aquellos datos extremos que, dada su distancia con respecto al grueso de la distribución, a priori resultan no ser congruentes si los comparamos con la población/muestra analizada. Existen varias vías para detectarlos, no obstante el método más rápdio y común se basa en el uso de boxplots. Asimismo, en este trabajo también haremos uso de la función “boxplots.stats()” a fin de detectar, con un output más allá del visual, los outliers que encontramos usando boxplots. Dado que muchas de las variables que hemos generado en la fase de limpieza son dicotómicas, tan solo nos encargaremos de comprobar si existen valores extremos en las columnas de **num\_rooms**, **latitude**, **longitude**, **hotel\_score** y el resto de **scores del hotel**, **length\_description**, **min\_price** y **max\_price**.

# Graficamos box-plot con ggplot  
ggplot(booking, aes(y = num\_rooms, color = city))+  
 geom\_boxplot() +  
 facet\_grid(~city)

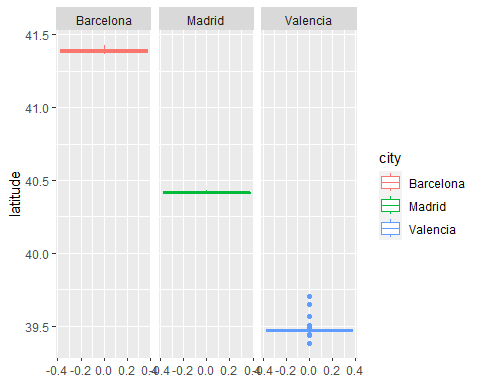


# Sacamos los valores considerados outliers  
boxplot.stats(booking$num\_rooms)$out

## integer(0)

Empezamos analizando el número de habitaciones y observamos que en las 3 ciudades el número oscila entre 2 y 3 habitaciones para las búsquedas realizadas, lo cual son números razonables, aunque en Valencia esta aparezca como un outlier. Por lo tanto, aceptamos los valores outliers observados en la ciudad de Valencia como valores correctos.

# Graficamos box-plot con ggplot  
ggplot(booking, aes(y = latitude, color = city))+  
 geom\_boxplot() +  
 facet\_grid(~city)

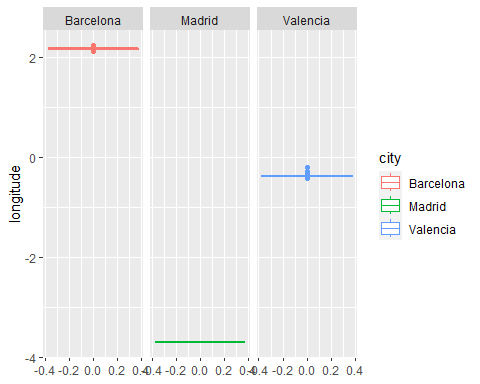


# Sacamos los valores considerados outliers  
boxplot.stats(booking$latitude)$out

## numeric(0)

Observamos que la latitud contiene valores distintos para cada ciudad, como es de esperar al tener cada una una ubicación distinta. En el caso de Valencia vemos una mayor dispersión y tenemos la presencia de algunos outliers, esto es debido a que probablemente, al ser una ciudad más pequeña que Madrid o Barcelona tenga una menor disponibilidad hotelera y algunos de los hoteles hayados no pertenezcan a Valencia ciudad sino a los alrededores de esta.

# Graficamos box-plot con ggplot  
ggplot(booking, aes(y = longitude, color = city))+  
 geom\_boxplot() +  
 facet\_grid(~city)

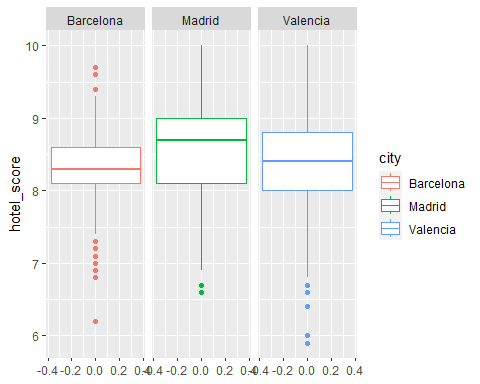


# Sacamos los valores considerados outliers  
boxplot.stats(booking$longitude)$out

## numeric(0)

Para la longitud obtenemos unos resultados parecidos, pues hay una menor dispersión en los outliers de Valencia. Dado que estas variaciones tan pequeñas en los resultados son razonables, aceptamos los resultados como validos.

# Graficamos box-plot con ggplot  
ggplot(booking, aes(y = hotel\_score, color = city))+  
 geom\_boxplot() +  
 facet\_grid(~city)



# Sacamos los valores considerados outliers  
boxplot.stats(booking$hotel\_score)$out

## [1] 6.9 6.9 6.9 6.9 6.9 10.0 6.6 6.6 5.9 7.0 6.6 6.0 6.6 5.9 6.7  
## [16] 10.0 5.9 7.0 6.8 6.6 6.7 6.9 6.0 6.9 6.4 10.0 6.6 6.6 10.0 7.0  
## [31] 6.9 6.9 6.9 10.0 7.0 6.6 7.0 6.6 6.7 7.0 6.4 7.0 6.9 6.8 10.0  
## [46] 6.9 6.9 7.0 6.2 6.8 7.0 6.6 6.7 7.0 6.9 6.9 6.9 6.9 7.0 7.0  
## [61] 6.9 6.9 7.0 5.9 6.0 7.0 6.9 7.0 6.8 6.8 5.9 7.0 10.0 6.6 6.6  
## [76] 6.6 7.0 6.9 10.0 6.6 10.0 10.0 6.6

En cuanto hotel\_score, por un lado observamos que la puntuación general del hotel oscila entre el 8 y el 9 siendo mayor en Madrid que en Barcelona o Valencia. También nos encontramos que hay outliers en las 3 ciudades no obstante todos se hayan en una puntuación entre el 5 y el 10, cuando el rango factible es entre 0 y 10, por lo que pueden ser valores anómalos de hoteles que han obtenido una puntuación más baja de lo normal, pero no por ello deben ser descartados. El análisis de esta variable ser presenta a modo de demostración general:

# Seleccionamos muestras de habitaciones con nota menor a 7  
kable(booking %>%  
 filter(hotel\_score <= 7.0) %>%  
 dplyr::select(c("hotel\_score", "facilities\_score", "cleanliness\_score", "location\_score")) %>%  
 sample\_n(5))

| hotel\_score | facilities\_score | cleanliness\_score | location\_score |
| --- | --- | --- | --- |
| 5.9 | 5.9 | 5.9 | 7.6 |
| 6.9 | 6.8 | 7.5 | 8.2 |
| 6.8 | 6.5 | 6.7 | 9.6 |
| 6.9 | 7.1 | 8.3 | 9.4 |
| 6.4 | 6.5 | 7.0 | 8.7 |

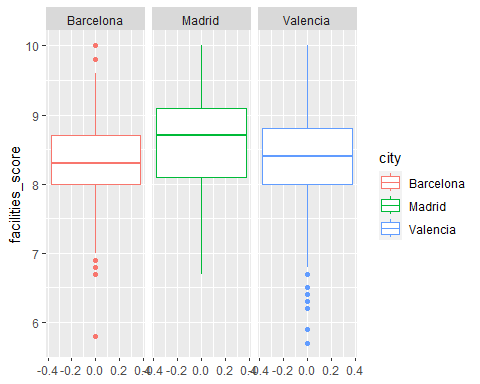
Analizando el resto de scores, se observa que los hoteles con una mala puntuación lo hacen en la mayoría de sus características, lo que corrobora los resultados obtenidos.

# Seleccionamos muestras de habitaciones con nota de 10  
kable(booking %>%  
 filter(hotel\_score == 10.0) %>%  
 dplyr::select(c("hotel\_score", "facilities\_score", "cleanliness\_score", "location\_score")) %>%  
 sample\_n(3))

| hotel\_score | facilities\_score | cleanliness\_score | location\_score |
| --- | --- | --- | --- |
| 10 | 10 | 10 | 10 |
| 10 | 10 | 10 | 10 |
| 10 | 10 | 10 | 10 |

Los hoteles con una puntuación de 10 destacan por obtener esta puntuación en todos los aspectos. Dado que solo corresponde a 3 registros y que los resultados son razonables, puede corresponder a un hotel nuevo con muy pocas valoraciones, lo que permitiría una nota excepcionalmente alta otorgada por unos pocos clientes.

# Graficamos box-plot con ggplot  
ggplot(booking, aes(y = facilities\_score, color = city))+  
 geom\_boxplot() +  
 facet\_grid(~city)



# Sacamos los valores considerados outliers  
boxplot.stats(booking$facilities\_score)$out

## [1] 5.9 6.5 5.7 6.4 5.9 5.9 6.5 6.4 6.3 5.9 6.2 6.5 6.5 5.8 6.5 5.9 5.7 6.5 6.5  
## [20] 5.9

Vemos que los resultados obtenidos continuan en la línea anterior con algunos outliers en Barcelona o Valencia, pero con resultados lógicos.

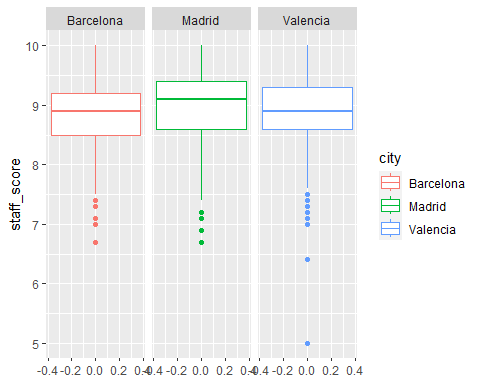
A continuación evaluamos aquellos hoteles que han obtenido una nota inferior al 6.5 para identificar si se trata de un error o no.

# Seleccionamos una muestra de hoteles con facilities\_score menor a 6.5  
kable(booking %>%  
 filter(facilities\_score <= 6.5) %>%  
 dplyr::select(c("hotel\_score", "facilities\_score", "cleanliness\_score", "location\_score")) %>%  
 sample\_n(3))

| hotel\_score | facilities\_score | cleanliness\_score | location\_score |
| --- | --- | --- | --- |
| 6.7 | 6.3 | 6.8 | 9.0 |
| 5.9 | 5.9 | 5.9 | 7.6 |
| 6.9 | 6.5 | 7.6 | 9.3 |

Observamos que los hoteles parecen tener unas características consistentes y que la extracción de características ha sido correcta por lo que no podemos atribuir los resultados a un error sino meramente a una baja calidad del hotel.

# Graficamos box-plot con ggplot  
ggplot(booking, aes(y = staff\_score, color = city))+  
 geom\_boxplot() +  
 facet\_grid(~city)



# Sacamos los valores considerados outliers  
boxplot.stats(booking$staff\_score)$out

## [1] 7.0 7.5 7.4 7.4 7.5 7.4 7.5 7.5 7.5 7.5 6.7 7.5 7.1 7.3 7.3 6.4 7.1 7.2  
## [19] 7.1 7.4 7.4 7.5 7.0 7.4 7.1 7.3 7.2 7.4 7.5 7.5 7.5 6.7 7.4 7.5 7.2 7.1  
## [37] 6.9 7.5 7.4 7.5 7.5 7.5 7.4 7.4 5.0 5.0 7.3 7.3 7.2 7.5 7.1 7.4 7.2 7.3  
## [55] 7.3 7.4 7.3 7.1 5.0 7.3 7.4 7.5 7.0 7.5 7.3 7.4 7.4 7.1 6.7 7.5 7.4 7.5  
## [73] 7.4 7.5 6.7 7.5 7.5 7.4 7.4 7.1 7.5 7.5 7.4 7.5 7.5 7.5 7.4 7.2 7.3 7.5  
## [91] 7.5 7.4 6.7 7.5 7.4 7.4 7.5 7.3 7.3 7.1 6.4 7.3 7.3 7.3 7.3 7.1 7.3 7.5  
## [109] 7.5 6.7 7.5 7.5 7.4 7.5 7.5 7.1 7.4 6.9 7.4 7.5 7.4 6.7 7.4 7.5 6.7 7.4  
## [127] 7.1 7.5 7.5 7.5 7.2

Para la puntuación otorgada a los trabajadores del hotel obtenemos unos resultados similares a los previos, con la diferencia de que en este apartado la puntuación promedio es más elevada rondando el 9 en las 3 ciudades y todas presentan outliers con valores inferiores.

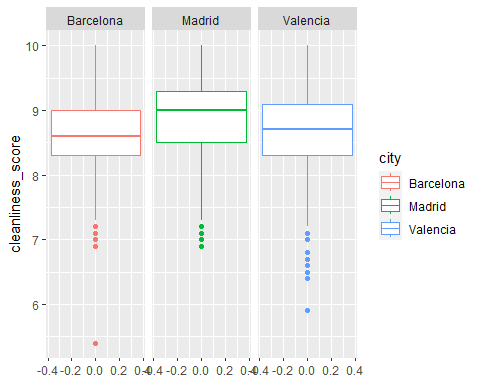
Evaluamos aquellos resultados con puntuaciones muy bajas.

# Seleccionamos una muestra de hoteles con staff\_score menor a 6.7  
kable(booking %>%  
 filter(staff\_score <= 6.7) %>%  
 dplyr::select(c("hotel\_score", "facilities\_score", "staff\_score", "location\_score")) %>%  
 sample\_n(3))

| hotel\_score | facilities\_score | staff\_score | location\_score |
| --- | --- | --- | --- |
| 7.3 | 7.0 | 6.7 | 9.4 |
| 7.3 | 7.0 | 6.7 | 9.4 |
| 6.0 | 5.7 | 6.4 | 8.9 |

De los resultados obtenidos sorprende uno en Valencia que tiene unas puntuaciones muy elevadas excepto en **staff\_score** la cual es bastante baja. Si nos fijamos estas puntuaciones son muy exactas (10.0, 9.0, 7.5 o 5.0) lo cual puede venir dado por un número de puntuaciones muy reducidas en las que un cliente ha puntuado de forma muy negativa este aspecto. No por ello eliminaremos el registro o lo modificaremos ya que es razonable el resultado, y afectará en la posición que presenta el hotel en el buscador.

# Graficamos box-plot con ggplot  
ggplot(booking, aes(y = cleanliness\_score, color = city))+  
 geom\_boxplot() +  
 facet\_grid(~city)



# Sacamos los valores considerados outliers  
boxplot.stats(booking$cleanliness\_score)$out

## [1] 7.2 6.9 7.2 7.0 7.0 6.9 7.1 5.9 6.8 7.0 7.1 6.4 7.1 5.9 7.2 6.7 7.1 7.1 7.2  
## [20] 5.9 6.4 6.7 7.1 6.8 6.6 6.5 7.0 6.9 7.1 7.0 7.1 7.1 7.0 6.9 6.8 7.0 6.7 7.1  
## [39] 7.0 6.8 7.0 6.8 7.1 7.1 7.2 7.2 6.9 6.9 5.4 7.1 7.2 7.0 6.9 6.8 7.2 6.5 7.2  
## [58] 6.9 6.8 5.9 6.4 6.8 6.5 7.1 6.8 6.7 7.1 6.4 5.9 6.8 7.2 6.9 6.9 7.0 7.1 7.1  
## [77] 7.0 7.0 7.0 7.2 7.1 6.9 6.9 7.0

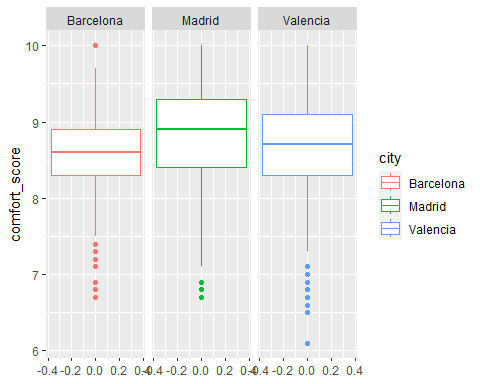
A continuación se evalúa la limpieza del hotel. Observamos unos resultados muy similares a los anteriores. Para descartar algún error observamos aquellos registros con puntuaciones excepcionalmente bajas.

# Seleccionamos una muestra de hoteles con cleanliness\_score menor a 7  
kable(booking %>%  
 filter(cleanliness\_score < 7.0) %>%  
 dplyr::select(c("hotel\_score", "facilities\_score", "cleanliness\_score", "location\_score")) %>%  
 sample\_n(3))

| hotel\_score | facilities\_score | cleanliness\_score | location\_score |
| --- | --- | --- | --- |
| 6.7 | 6.7 | 6.7 | 8.2 |
| 6.7 | 6.3 | 6.8 | 9.0 |
| 7.0 | 7.4 | 6.8 | 6.8 |

Vemos un comportamiento razonable en estos hoteles, por lo que, siendo resultados razonables, aceptamos estos valores outliers. Se ha seguido el mismo proceso con todo el resto de scores, alcanzando las mismas conclusiones, por lo que no eliminaremos ningún registro que presenta valores “atípicos” en columnas de scores pues sus valores los consideramos lógicos.

# Graficamos box-plot con ggplot  
ggplot(booking, aes(y = comfort\_score, color = city))+  
 geom\_boxplot() +  
 facet\_grid(~city)



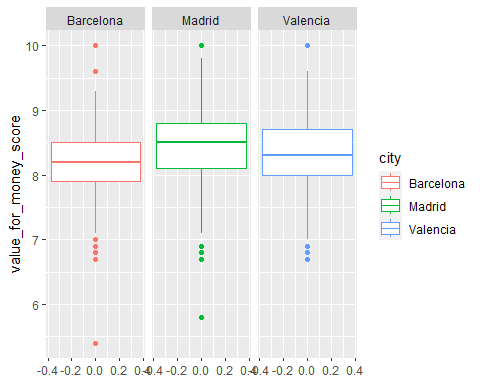
# Sacamos los valores considerados outliers  
boxplot.stats(booking$comfort\_score)$out

## [1] 7.1 7.1 7.1 6.8 6.9 7.1 6.7 6.1 7.1 6.8 6.1 6.7 6.1 7.0 7.1 7.1 6.1 7.0 6.7  
## [20] 6.7 6.6 6.9 6.5 6.6 6.9 7.1 6.7 7.1 7.1 6.9 6.8 7.1 6.9 6.8 7.0 7.1 7.1 6.8  
## [39] 6.7 7.1 6.7 7.1 7.1 6.9 7.1 7.1 7.0 7.1 7.1 7.1 7.1 6.7 7.1 6.1 6.1 7.1 7.0  
## [58] 7.1 7.1 6.7 6.7 6.1 6.9 7.1 6.9 6.7 6.8 7.1 6.9 6.9

# Seleccionamos una muestra de hoteles con comfort\_score menor a 7  
kable(booking %>%  
 filter(comfort\_score < 7.0) %>%  
 dplyr::select(c("hotel\_score", "facilities\_score", "comfort\_score", "location\_score")) %>%  
 sample\_n(3))

| hotel\_score | facilities\_score | comfort\_score | location\_score |
| --- | --- | --- | --- |
| 7.1 | 7.0 | 6.9 | 9.1 |
| 6.9 | 7.1 | 6.8 | 9.4 |
| 5.9 | 5.9 | 6.1 | 7.6 |

# Graficamos box-plot con ggplot  
ggplot(booking, aes(y = value\_for\_money\_score, color = city))+  
 geom\_boxplot() +  
 facet\_grid(~city)



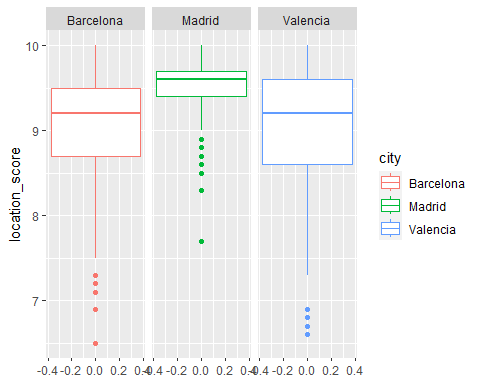
# Sacamos los valores considerados outliers  
boxplot.stats(booking$value\_for\_money\_score)$out

## [1] 6.9 6.9 10.0 10.0 10.0 5.8 6.7 10.0 6.7 10.0 10.0 6.7 6.9 10.0 6.7  
## [16] 6.8 6.9 6.7 6.7 6.9 10.0 10.0 10.0 10.0 5.8 10.0 10.0 9.8 10.0 10.0  
## [31] 10.0 6.8 9.8 10.0 10.0 10.0 10.0 10.0 6.9 10.0 9.8 10.0 10.0 6.7 6.9  
## [46] 10.0 10.0 6.9 6.9 10.0 6.7 6.7 5.4 6.9 10.0 10.0 6.8 10.0 10.0 10.0  
## [61] 6.9 6.8 10.0 10.0 6.7 6.7 6.7 10.0 6.9 6.7 10.0 10.0 10.0 5.8 6.9  
## [76] 10.0 10.0 9.8 10.0 10.0 10.0

# Seleccionamos una muestra de hoteles con value\_for\_money\_score menor a 7  
kable(booking %>%  
 filter(value\_for\_money\_score <= 7.0) %>%  
 dplyr::select(c("hotel\_score", "facilities\_score", "value\_for\_money\_score", "location\_score")) %>%  
 sample\_n(3))

| hotel\_score | facilities\_score | value\_for\_money\_score | location\_score |
| --- | --- | --- | --- |
| 6.6 | 6.7 | 5.8 | 8.8 |
| 6.9 | 6.8 | 7.0 | 8.2 |
| 6.9 | 6.8 | 7.0 | 8.2 |

# Graficamos box-plot con ggplot  
ggplot(booking, aes(y = location\_score, color = city))+  
 geom\_boxplot() +  
 facet\_grid(~city)



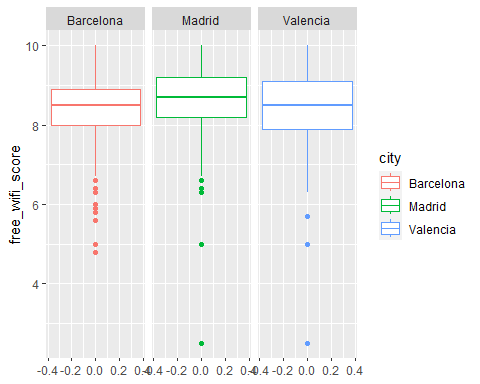
# Sacamos los valores considerados outliers  
boxplot.stats(booking$location\_score)$out

## [1] 7.6 7.3 7.6 7.6 7.2 7.6 6.8 6.7 7.6 7.5 7.7 7.3 7.8 7.6 7.8 7.5 7.8 7.8  
## [19] 7.6 6.8 7.8 6.8 7.8 7.5 7.7 6.6 7.3 7.8 7.5 7.8 7.5 7.8 7.5 7.3 7.6 7.6  
## [37] 7.6 7.7 6.8 7.5 7.5 6.7 7.6 7.6 7.3 7.3 7.5 6.8 6.8 6.8 7.5 7.7 7.7 6.9  
## [55] 7.7 7.8 6.8 7.5 7.8 7.3 7.6 7.5 7.2 7.3 7.8 7.1 7.8 7.2 6.5 6.9 7.6 6.8  
## [73] 7.7 6.8 6.7 7.6 7.6 7.3 7.6 7.6 7.6 7.8 7.5 6.9 7.2 7.2 6.8 7.6 7.8 7.8  
## [91] 6.8 7.7 7.5 7.5 7.8 7.7 7.5 6.8 6.8 6.8 7.8 7.8 7.8 7.7 7.6 7.8 7.7

# Seleccionamos una muestra de hoteles con location\_score menor a 7  
kable(booking %>%  
 filter(location\_score < 7.0) %>%  
 dplyr::select(c("hotel\_score", "facilities\_score", "value\_for\_money\_score", "location\_score")) %>%  
 sample\_n(3))

| hotel\_score | facilities\_score | value\_for\_money\_score | location\_score |
| --- | --- | --- | --- |
| 7.9 | 8.0 | 8.1 | 6.8 |
| 7.9 | 8.0 | 7.7 | 6.9 |
| 8.6 | 9.2 | 9.1 | 6.7 |

# Graficamos box-plot con ggplot  
ggplot(booking, aes(y = free\_wifi\_score, color = city))+  
 geom\_boxplot() +  
 facet\_grid(~city)



# Sacamos los valores considerados outliers  
boxplot.stats(booking$free\_wifi\_score)$out

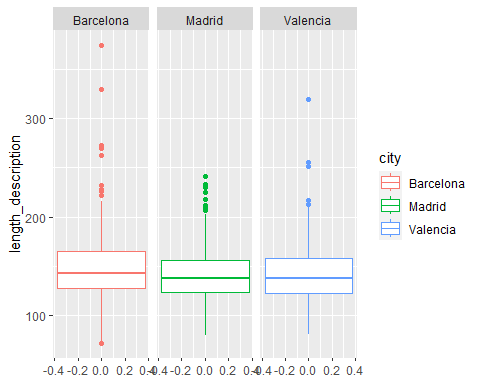
## [1] 5.9 5.6 6.6 5.9 5.6 6.4 6.3 6.3 6.6 5.0 6.6 6.5 5.0 5.0 6.5 6.3 5.0 6.5  
## [19] 6.3 2.5 6.7 5.0 6.3 6.6 4.8 5.9 5.6 6.6 5.9 5.6 6.6 4.8 5.0 2.5 6.4 5.9  
## [37] 5.6 6.6 4.8 6.0 5.6 6.4 5.0 6.7 5.0 5.0 5.7 6.4 6.5 5.0 5.0 5.0 5.6 6.4  
## [55] 6.4 6.6 5.0 5.9 6.3 5.6 6.7 5.8 6.6 5.0 6.3 5.0 6.3 5.0 5.9 5.6 6.6 4.8  
## [73] 5.6 5.6 5.9 5.6 4.8 5.9 5.9 6.3 5.9 6.6 6.4 5.9 5.6 5.6 5.8 6.7 5.9 4.8  
## [91] 5.9 5.0 6.6 5.6 5.0 6.3 5.0 5.0 6.5 6.5 5.0 6.5 5.0 5.7 6.5 5.0 6.6 6.6  
## [109] 6.3 5.0 6.6

# Seleccionamos una muestra de hoteles con free\_wifi\_score menor a 5  
kable(booking %>%  
 filter(free\_wifi\_score < 5.0 & free\_wifi\_score > 0) %>%  
 dplyr::select(c("hotel\_score", "facilities\_score", "free\_wifi\_score", "location\_score")) %>%  
 sample\_n(3))

| hotel\_score | facilities\_score | free\_wifi\_score | location\_score |
| --- | --- | --- | --- |
| 7.1 | 6.7 | 4.8 | 9.2 |
| 7.8 | 7.9 | 2.5 | 9.6 |
| 7.1 | 6.7 | 4.8 | 9.1 |

A continuación se evalúa la longitud de las descripciones.

# Graficamos box-plot con ggplot  
ggplot(booking, aes(y = length\_description, color = city))+  
 geom\_boxplot() +  
 facet\_grid(~city)



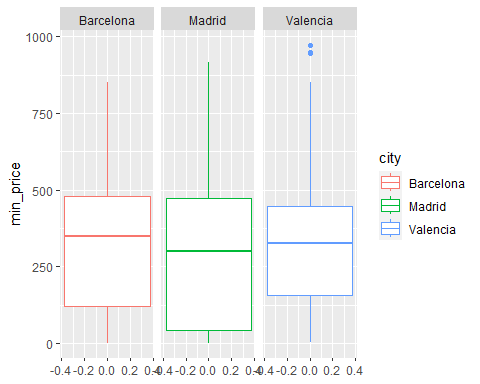
# Sacamos los valores considerados outliers  
boxplot.stats(booking$length\_description)$out

## [1] 226 232 215 270 226 232 270 374 215 226 232 270 330 215 273 228 222 217 213  
## [20] 320 251 256 222 226 232 72 216 215 226 218 212 232 225 263 222 233 230 212  
## [39] 232 270 330 222 263 73 212 212 218 232 212 241 217 215 226 232 232 228 215  
## [58] 330 216 222 226 72 270 232 222 222 222 222 270 222 226 228 330 263 73 251  
## [77] 213 320 212 218 230 241 212

Observamos que las descripciones tienen una longitud muy similar en las 3 ciudades, en torno a las 120-170 palabras, presentando algún outlier por la parte superior, pero siempre con descripciones inferiores a las 400 palabras. Aunque corresponda a una descripción un poco extensa, no parece que corresponda a un error sino simplemente a una descripción más extensa de lo habitual.

A continuación evaluaremos el precio mínimo de las habitaciones del hotel.

# Graficamos box-plot con ggplot  
ggplot(booking, aes(y = min\_price, color = city))+  
 geom\_boxplot() +  
 facet\_grid(~city)

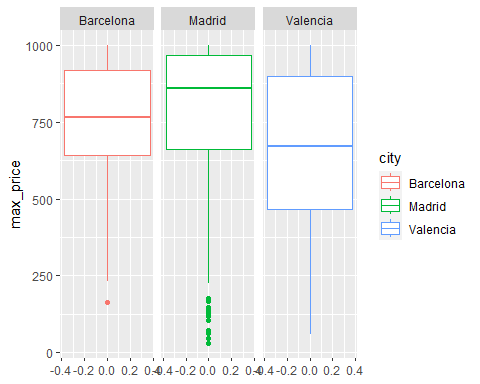


# Sacamos los valores considerados outliers  
boxplot.stats(booking$min\_price)$out

## numeric(0)

Observamos que el precio mínimo tiene una gran disparidad oscilando de promedio entre los 100€ y los 500€, siendo Valencia la ciudad con el menor rango intercuartil con presencia de outliers en el rango superior de precios, con alguna habitación con precios cercanos a los 1000€, que aunque puedan parecer caros, son razonables en algún hotel de lujo de la ciudad.

# Graficamos box-plot con ggplot  
ggplot(booking, aes(y = max\_price, color = city))+  
 geom\_boxplot() +  
 facet\_grid(~city)



# Sacamos los valores considerados outliers  
boxplot.stats(booking$max\_price)$out

## [1] 30 47 30 47

A continuación se evalúan los precios máximos. Observamos unos precios promedio mucho más elevados con un mayor rango intercuartílico en la ciudad de Valencia, lo que sorprende al ser lo contrario que con los precios mínimos. También se observa que los outliers en este caso aparecen por el lado inferior especialmente en la ciudad de Madrid, aunque con precios positivos, que podrían corresponder a hoteles muy económicos o albergues.

Por otro lado sorprende ver que los precios máximos no superan el valor máximo observado en el precio mínimo, lo que se puede atribuir a que aquellos hoteles con precios muy elevados no ofrecían una variedad de habitaciones y precios.

## 3. INTEGRACIÓN Y SELECCIÓN

La sección de integración y selección la encontramos tras la de limpieza, pues muchas de las columnas que queriamos descartar han sido tratadas en la fase de limpieza para extraer información de utilidad. Hecho esto, en la presente sección procedemos a eliminar todas las columnas del dataset original que ya no vamos a necesitar para los siguientes apartados.

# Seleccionamos con la función "select" de dplyr todas las columnas excepto las que ya hemos limpiado para crear otras, además de aquellasque nos van a ser de poca utilidad.  
booking <- booking %>%  
 dplyr::select(-c("name", "search\_date", "hotel\_coordinates", "hotel\_scores", "check.in", "check.out", "address", "features", "hotel\_description", "room\_data"))

Por otro lado, hemos optado 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 o los precios de dichos alojamientos.
* 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 y, a su vez, directamente en los precios de los hoteles.

# Leemos los datos de vuelos  
flights <- read.csv("../data/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

Para añadir dicha información hacemos un tratamiento de los datos obtenidos en Eurostat a fin de obtener una media de vuelos por mes para los últimos 5 años. Una vez obtenidos, tan solo debemos hacer un merge con nuestro dataset original.

proc\_flights <- flights %>%   
 # Filtramos para quedarnos solo con los aeropuertos de interés  
 # Código OACI/ICAO: Barcelona --> ES\_LEBL; Valencia --> ES\_LEVC; Madrid: ES\_LEMD  
 filter(rep\_airp %in% c("ES\_LEBL", "ES\_LEVC", "ES\_LEMD")) %>%   
 # Filtramos para quedarnos solo con los datos de carga de pasajeros  
 filter(tra\_meas == "PAS\_CRD") %>%  
 # Nos interesa solo los datos mensuales, así que filtramos por ellos  
 filter(freq == "M") %>%  
 separate(TIME\_PERIOD, c("YEAR", "MONTH"), sep = "-") %>%  
 # Nos quedamos solo con los últimos 5 años y los meses de Marzo, Junio y Diciembre  
 filter(YEAR %in% c("2022", "2021", "2020", "2019", "2018") & MONTH %in% c("03", "06", "12")) %>%  
 # No nos interesa los datos por aerolínea, solo los generales  
 filter(airline == "TOTAL") %>%  
 # Modificamos la columna de MONTH para tener los nombres del mes y cambiamos el código del aeropuerto por el nombre de la ciudad  
 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"))) %>%  
 # Agrupamos por ciudad y mes y calculamos la media de vuelos.  
 group\_by(rep\_airp, month\_name) %>%  
 mutate(mean\_flights = mean(OBS\_VALUE)) %>%  
 ungroup() %>%  
 dplyr::select(c(city\_airp, month\_name, mean\_flights))  
  
# Nos quedamos con los registros únicos  
final\_flights <- unique(proc\_flights)  
  
# Hacemos un merge con bookings  
booking <- booking %>%  
 merge(final\_flights, by.x = c("city", "month"), by.y = c("city\_airp", "month\_name"))  
  
# Exportamos los datos a un archivo csv  
write.csv(booking, "../data/hotel\_data\_processed.csv")

## 4. ANÁLISIS DE LOS DATOS

A continuación, se evaluará la influencia de las distintas variables sobre la posición de un hotel en el buscador y sobre el precio que este fija para sus habitaciones. Para ello se empleará una matriz de correlaciones. Una vez obtenidas las variables con mayor correlación con la variable objetivo generaremos una selección de los atributos a analizar, un grupo de variables distinto por cada problema a tratar.

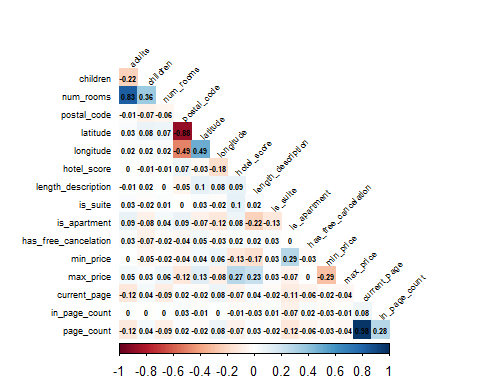
Tras generar los distintos grupos procederemos a evaluar la normalidad y la homogeneidad de la varianza para así poder determinar en el siguiente punto que pruebas estadísticas se pueden aplicar.

Finalmente aplicaremos distintos tests estadísticos adicionales a la correlación hallada inicialmente. Estas pruebas estadísticas incluirán contrastes de hipótesis y regresiones lineales (enfocados desde un punto de vista inferencial, no predictivo).

#### 4.1 Correlación con la variable objetivo y selección de los grupos de datos que se quieren analizar

Comenzamos evaluando la correlación (de Spearman, dado que no podemos asumir normalidad y utilizamos variables categóricas también) entre las distintas variables con nuestra variable objetivo, para ello creamos varias matrices de correlación y evaluamos los resultados obtenidos.

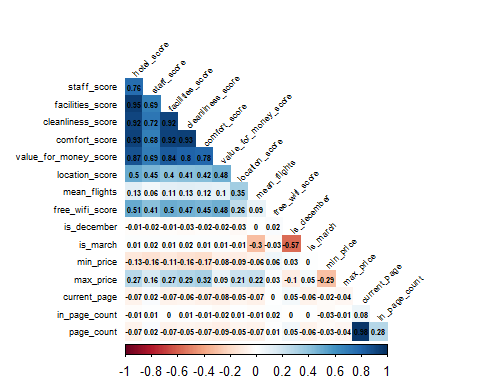
# Transformamos las variables a numérico  
booking$postal\_code <- as.numeric(booking$postal\_code)  
booking$latitude <- as.numeric(booking$latitude)  
booking$longitude <- as.numeric(booking$longitude)  
booking$free\_wifi <- as.numeric(booking$free\_wifi)  
booking$is\_apartment <- as.numeric(booking$is\_apartment)  
booking$pet\_friendly <- as.numeric(booking$pet\_friendly)  
booking$page\_count <- booking$page\_count + 1  
  
# Seleccionamos solo las variables numéricas  
booking\_numeric <- booking %>%  
 dplyr::select(-c("city", "month", "as.integer(!is.na(NA))"))  
  
mattmp <- booking\_numeric %>%  
 dplyr::select(c("adults", "children", "num\_rooms", "postal\_code", "latitude", "longitude", "hotel\_score", "length\_description",   
 "is\_suite", "is\_apartment", "has\_free\_cancelation", "min\_price","max\_price", "current\_page", "in\_page\_count", "page\_count"))  
  
# Creamos una matriz de correlación del dataset  
cor\_mat <- cor(mattmp, method = "spearman")  
plot.new()  
plot.window(xlim=c(-2,2), ylim=c(5,10))  
corrplot(cor\_mat, method="color", type = "lower", tl.col = "black", tl.srt = 45, diag = FALSE, tl.cex = 0.55, addCoef.col = 'black', number.cex = 0.5)



mattmp <- booking\_numeric %>%  
 dplyr::select(c("hotel\_score", "staff\_score", "facilities\_score", "cleanliness\_score", "comfort\_score", "value\_for\_money\_score", "location\_score",  
 "mean\_flights", "free\_wifi\_score", "is\_december", "is\_march", "min\_price", "max\_price", "current\_page", "in\_page\_count", "page\_count"))  
  
sapply(mattmp, function(y) sum(is.na(y)))

## hotel\_score staff\_score facilities\_score   
## 0 0 0   
## cleanliness\_score comfort\_score value\_for\_money\_score   
## 0 0 0   
## location\_score mean\_flights free\_wifi\_score   
## 0 0 0   
## is\_december is\_march min\_price   
## 0 0 0   
## max\_price current\_page in\_page\_count   
## 0 0 0   
## page\_count   
## 0

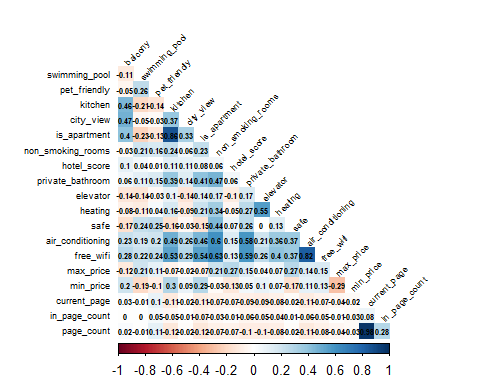
# Creamos una matriz de correlación del dataset  
cor\_mat <- cor(mattmp, method = "spearman")  
plot.new()  
plot.window(xlim=c(-2,2), ylim=c(5,10))  
corrplot(cor\_mat, method="color", type = "lower", tl.col = "black", tl.srt = 45, diag = FALSE, tl.cex = 0.55, addCoef.col = 'black', number.cex = 0.5)



mattmp <- booking\_numeric %>%  
 dplyr::select(c("balcony", "swimming\_pool", "pet\_friendly", "kitchen", "city\_view", "is\_apartment", "non\_smoking\_rooms", "hotel\_score",   
 "private\_bathroom", "elevator", "heating", "safe", "air\_conditioning", "free\_wifi", "max\_price", "min\_price", "current\_page", "in\_page\_count", "page\_count"))  
  
sapply(mattmp, function(y) sum(is.na(y)))

## balcony swimming\_pool pet\_friendly kitchen   
## 0 0 0 0   
## city\_view is\_apartment non\_smoking\_rooms hotel\_score   
## 0 0 0 0   
## private\_bathroom elevator heating safe   
## 0 0 0 0   
## air\_conditioning free\_wifi max\_price min\_price   
## 0 0 0 0   
## current\_page in\_page\_count page\_count   
## 0 0 0

# Creamos una matriz de correlación del dataset  
cor\_mat <- cor(mattmp, method = "spearman")  
plot.new()  
plot.window(xlim=c(-2,2), ylim=c(5,10))  
corrplot(cor\_mat, method="color", type = "lower", tl.col = "black", tl.srt = 45, diag = FALSE, tl.cex = 0.55, addCoef.col = 'black', number.cex = 0.5)



Como bien se puede observar en las matrices, las variables que recogen las distintas puntuaciones están correlacionadas entre ellas, por lo que un hotel que haya obtenido una puntuación elevada en un aspecto generalmente también lo obtendrá en el resto. Este factor es de especial interés si se seleccionan varias columnas de este tipo, pues corremos el riesgo de sufrir un problema de multicolinealidad. También encontramos que los hoteles con distintos atributos como aire acondicionado o baño privado suelen tener el resto de servicios adicionales, aunque este factor puede deberse a que es más habitual encontrar este tipo de elementos en una habitación de hotel, por lo que es común que cuando exista una habitación con calefacción, por ejemplo, también tenga aire acondicionado o baño privado. Por último, aunque es evidente, cabe destacar que las mayores correlaciones las tenemos con variables directamente relacionadas o que presentan información de la misma fuente (current\_page con page\_count o max\_price con min\_price). No obstante, esto no es un problema, pues eliminaremos estas variables en los ajustes para no incurrir en “data leakage”.

# Filtramos para obtener cuales son las variables más correlacionadas con nuestra variable objetivo, current\_page  
cor\_mat <- cor(booking\_numeric, method = "spearman")

## Warning in cor(booking\_numeric, method = "spearman"): the standard deviation is  
## zero

cor\_mat\_df <- as.data.frame(cor\_mat)  
cor\_mat\_df %>%  
 filter(abs(page\_count) > 0.065) %>%  
 filter(page\_count < 0.95) %>%  
 arrange(page\_count) %>%  
 dplyr::select(page\_count)

## page\_count  
## kitchen -0.12023093  
## is\_apartment -0.12015885  
## adults -0.11580745  
## air\_conditioning -0.11400432  
## elevator -0.09860877  
## private\_bathroom -0.09661382  
## num\_rooms -0.08662662  
## value\_for\_money\_score -0.08574257  
## heating -0.08399204  
## free\_wifi -0.07529422  
## non\_smoking\_rooms -0.07424905  
## mean\_flights -0.06952514  
## comfort\_score -0.06936356  
## hotel\_score -0.06765554  
## facilities\_score -0.06508687  
## longitude 0.07601044  
## pet\_friendly 0.10535168  
## in\_page\_count 0.28025901

Cuando evaluamos la posición en la página web observamos que no hay una correlación elevada con ninguna variable, encontrando valores por debajo de 0.12 (en valor absoluto) para todos los casos. No obstante, para el caso de page\_count, hemos optado por someter a estudio aquellas variables que presentan mayor correlación de spearman con la variable dependiente (pues evaluando con distintas métricas que buscan determinar el nivel de dependencia no-lineal hemos obtenido resultados similares). Las variables seleccionadas en cuestión son: kitchen, is\_apartment, adults, air\_conditioning, elevator, private\_bathroom, num\_rooms, value\_for\_money\_score, heating, free\_wifi, non\_smoking\_rooms, comfort\_score, hotel\_score, longitude, facilities\_score pet\_friendly, mientras que la variable dependiente será page\_count.

En el caso del problema a resolver relacionado con el precio, al analizar las correlaciones sí que encontramos valores entre 0.15 y 0.20. Aunque estas correlaciones no sean muy elevadas nos ofrecen un escenario más prometedor que el anteriormente presentado. En este caso, las variables más relevantes son:

* **postal\_code**
* **balcony**
* **pet\_friendly**
* **elevator**
* **heating**
* **cleanliness\_score**
* **latitude**
* **mean\_flights**
* **location\_score**
* **hotel\_score**
* **facilities\_score**
* **non\_smoking\_rooms**
* **comfort\_score**
* **is\_apartment**
* **private\_bathroom**
* **kitchen**
* **air\_conditioning**

# Filtramos para obtener cuales son las variables más correlacionadas con nuestra variable objetivo, page\_count  
cor\_mat <- cor(booking\_numeric, method = "spearman")

## Warning in cor(booking\_numeric, method = "spearman"): the standard deviation is  
## zero

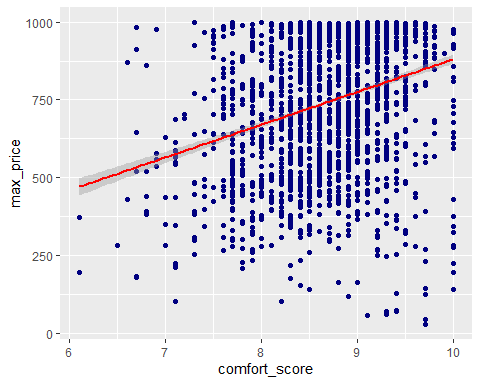
cor\_mat\_df <- as.data.frame(cor\_mat)  
cor\_mat\_df %>%  
 filter(abs(max\_price) > 0.1) %>%  
 filter(max\_price < 0.95) %>%  
 arrange(max\_price) %>%  
 dplyr::select(max\_price)

## max\_price  
## min\_price -0.2933215  
## postal\_code -0.1244939  
## balcony -0.1185999  
## pet\_friendly 0.1073362  
## latitude 0.1271236  
## air\_conditioning 0.1408840  
## private\_bathroom 0.1452074  
## free\_wifi 0.1523803  
## staff\_score 0.1640168  
## swimming\_pool 0.2055779  
## non\_smoking\_rooms 0.2110602  
## location\_score 0.2116932  
## mean\_flights 0.2224695  
## length\_description 0.2287604  
## facilities\_score 0.2667483  
## hotel\_score 0.2700575  
## safe 0.2719657  
## cleanliness\_score 0.2919506  
## comfort\_score 0.3206842

A modo ilustrativo podemos graficar un scatter plot para visualizar la variable que más parece estar relacionada con el precio, el comfort\_score.

ggplot(booking, aes(comfort\_score, max\_price)) +  
 geom\_point(color = "navy") +  
 geom\_smooth(method = "lm", color = "red")

## `geom\_smooth()` using formula = 'y ~ x'



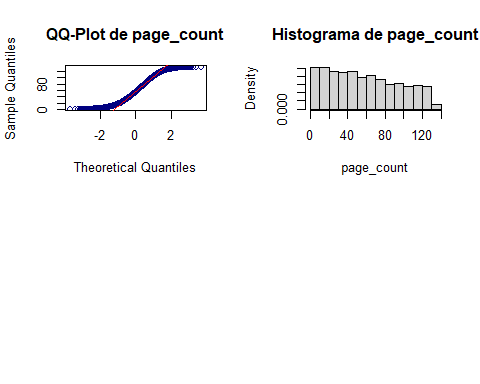
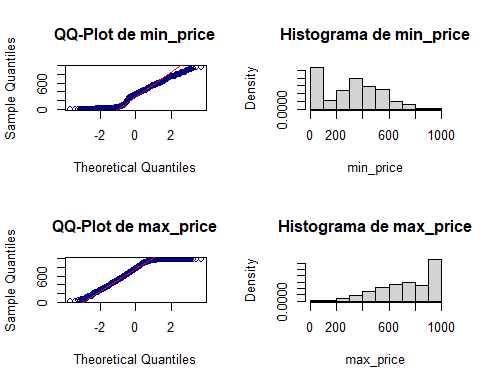
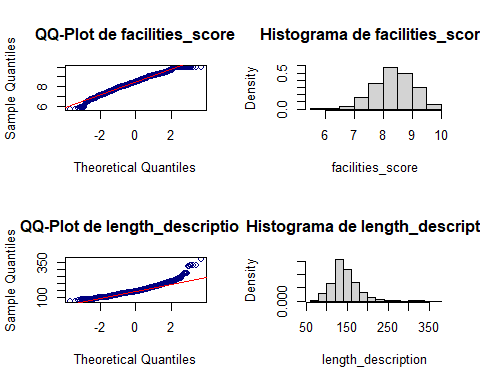
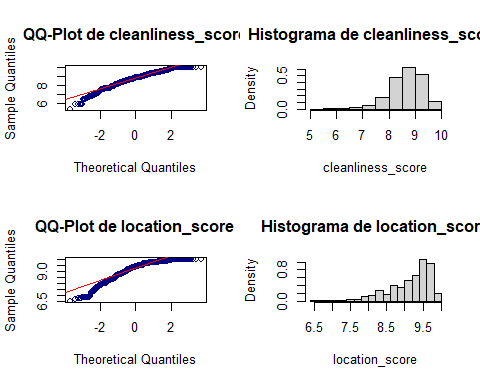
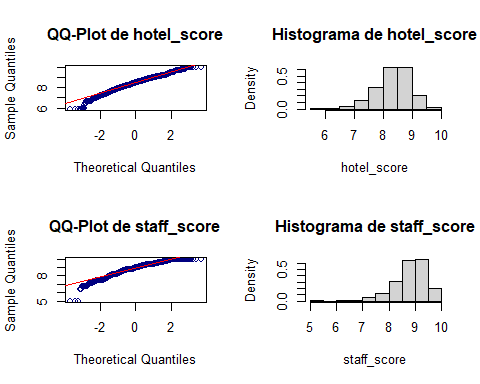
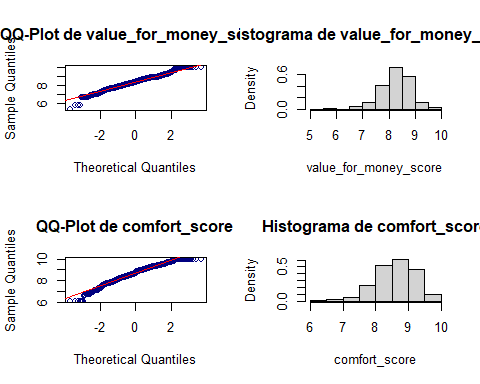
#### 4.2 Comprobación de la normalidad y homogeneidad de la varianza

Para comprobar si los valores de las variables de nuestro dataset se distribuyen o se aproximan a una población distribuida normalmente haremos un análisis visual, con QQ-Plots e Histogramas, y un análisis mediante el test Shapiro-Wilk. Concretamente, solo comprobaremos la normalidad para las variables seleccionadas en el apartado anterior.

# Seleccionamos las variables a analizar  
norm\_cols <- c("value\_for\_money\_score", "comfort\_score", "hotel\_score", "staff\_score", "cleanliness\_score", "location\_score", "facilities\_score", "length\_description", "min\_price", "max\_price", "page\_count")  
  
booking\_norm <- booking\_numeric %>%  
 dplyr::select(norm\_cols)

## Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.  
## i Please use `all\_of()` or `any\_of()` instead.  
## # Was:  
## data %>% select(norm\_cols)  
##   
## # Now:  
## data %>% select(all\_of(norm\_cols))  
##   
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.

# Ploteamos los qq-plots y los histogramas  
par(mfrow=c(2,2))  
for(i in 1:ncol(booking\_norm)){  
 qqnorm(booking\_norm[,i], main = paste("QQ-Plot de",colnames(booking\_norm)[i]), col = "navy")  
 qqline(booking\_norm[,i], col = "red")  
 hist(booking\_norm[,i], main = paste("Histograma de", colnames(booking\_norm)[i]), xlab=colnames(booking\_norm)[i], freq = FALSE)  
}



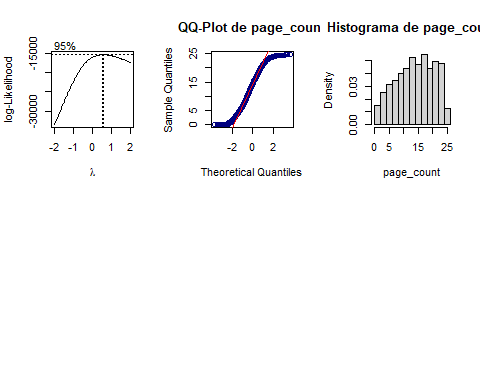
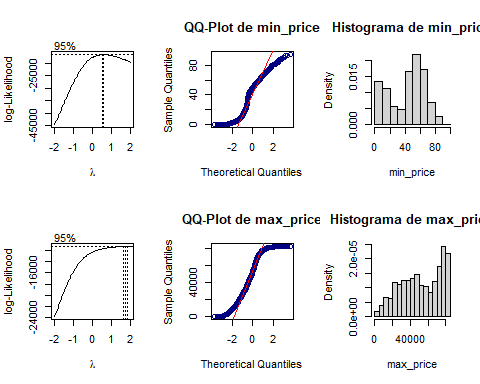
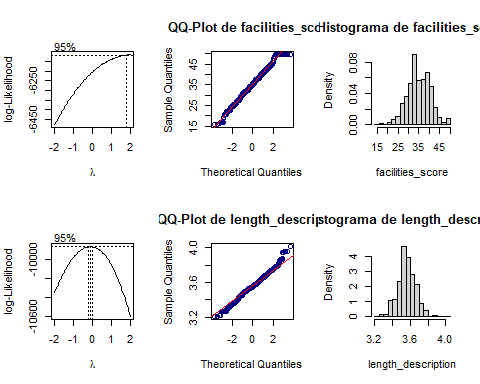
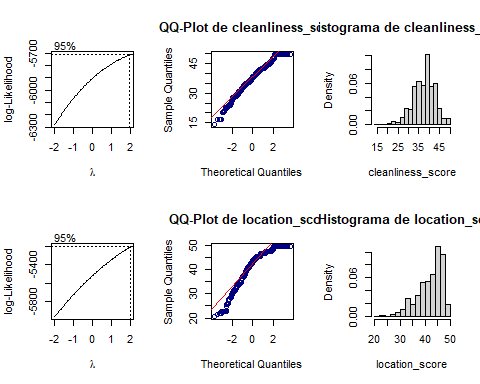
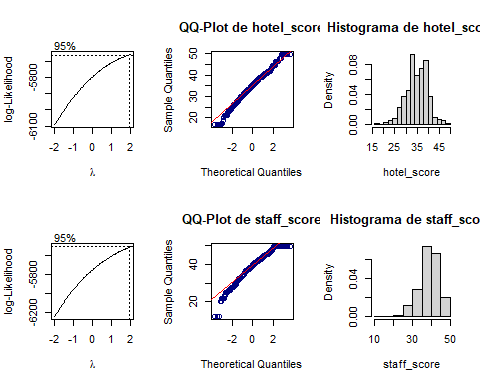
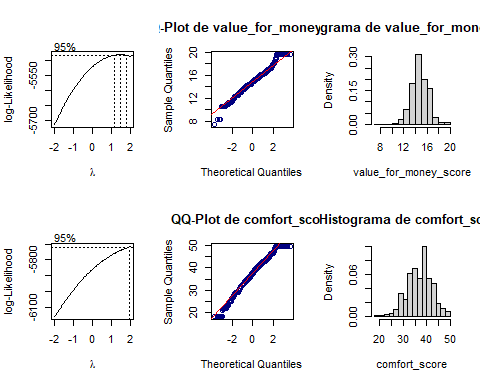
Los QQ-plots y los histogramas nos muestran una distribuciones con una evidente asimetría negativa en la mayoría de los casos (excepto variables como length\_description que muestran una asimetría más bien positiva, aunque podría ser influencia de un dato extremo que no ha sido considerado outlier), así como un evidente problema con las colas, pues en la mayoría de los casos los puntos se alejan de la línea roja, la cual corresponde a la hipotética distribución normal. Por tanto, a priori podríamos rechazar la hipótesis de normalidad de las variables. Es más, tras ejecutar el test Shapiro-Wilk confirmamos de forma más fiable nuestras primeras conclusiones, pues todos los p-valores muestran valores muy inferiores a 0.05, rechazando en todos los casos la hipótesis nula sobre la normalidad de los datos.

# Calculamos el test de Shapiro-Wilk para cada variable y sacamos su pvalor  
pvalues <- t(data.frame(lapply(booking\_norm, function(y) format(shapiro.test(y)$p.value, digits = 3))))  
  
kable(pvalues, col.names = c("p-value"))

|  | p-value |
| --- | --- |
| value\_for\_money\_score | 1.23e-18 |
| comfort\_score | 9.62e-22 |
| hotel\_score | 8.63e-21 |
| staff\_score | 8.08e-32 |
| cleanliness\_score | 2.19e-26 |
| location\_score | 8.67e-46 |
| facilities\_score | 1.54e-15 |
| length\_description | 6.06e-33 |
| min\_price | 6.87e-35 |
| max\_price | 3.71e-40 |
| page\_count | 2.73e-34 |

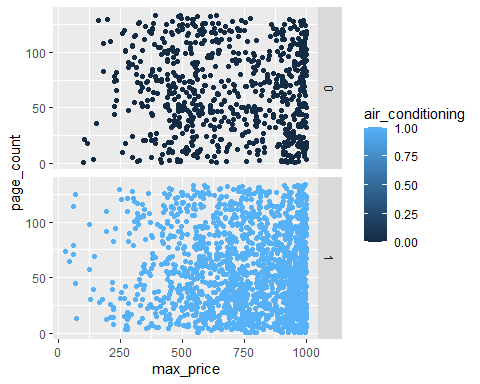
A fin de aproximar más los datos a una distribución normal, se propone hacer una transformación de box-cox y comprobar si los datos efectivamente se aproximan más a una normal. Tras llevar a cabo este proceso, si bien los histogramas muestran en algunos casos una distribución “visualmente más normales” y los QQ-plots parecen mostrar que la transformación ha aliviado ligeramente el problema con las colas, los resultados de la transformación no han sido suficientes como para poder considerar que los datos se distribuyen como una normal. No obstante, no debemos preocuparnos demasiado por este problema, pues tenemos suficientes registros como para poder apoyarnos en el Teorema del Límite Central. Por ejemplo, el t-test asume que las medias de las diferentes muestras se distribuyen normalmente (siendo una muestra de más de 30 registros suficiente, como rule of thumb, para que el t-test sea válido aunque los datos no se distribuyan normalmente).

# Ploteamos los qq-plots y los histogramas  
par(mfrow=c(2,3))  
for(i in 1:ncol(booking\_norm)){  
 b <- boxcox(lm(booking\_norm[,i] ~ 1))  
 lambda <- b$x[which.max(b$y)]  
 qqnorm((booking\_norm[,i]^lambda - 1)/lambda, main = paste("QQ-Plot de",colnames(booking\_norm)[i]), col = "navy")  
 qqline((booking\_norm[,i]^lambda - 1)/lambda, col = "red")  
 hist((booking\_norm[,i]^lambda - 1)/lambda, main = paste("Histograma de", colnames(booking\_norm)[i]), xlab=colnames(booking\_norm)[i], freq = FALSE)  
}



Seguidamente, analizaremos la homocedasticidad mediante scatterplots y varios test de homocedasticidad, o de homogeneidad de varianzas. En este caso, no podemos utilizar el test de Barlett ya que es muy sensible ante datos no normales y, como acabamos de presentar, los datos analizados no se aproximan en ningún caso a una normal. Por tanto, en este caso haremos uso del test de Fligner-Killeen.

# Generamos el scatter para compromar la homocedasticidad de page\_count cuando dividimos por air\_conditioning.  
ggplot(booking, aes(x = max\_price, y = page\_count, color = air\_conditioning)) +   
 geom\_point() +  
 facet\_grid(air\_conditioning~.)



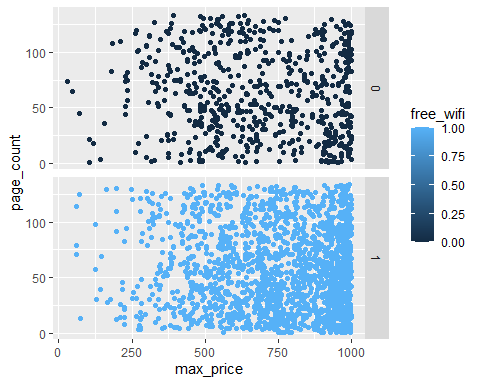
Como todos los test de homogeneidad de varianzas, nos basamos en las siguientes hipótesis:

Donde rechazaremos la hipótesis nula cuando el pvalor del test sea inferior a un nivel de significación igual a 0.05.

# Ajustamos el test de fligner-killeen para page\_count y air\_conditioning  
fligner.test(page\_count ~ air\_conditioning, data = booking)

##   
## Fligner-Killeen test of homogeneity of variances  
##   
## data: page\_count by air\_conditioning  
## Fligner-Killeen:med chi-squared = 0.20066, df = 1, p-value = 0.6542

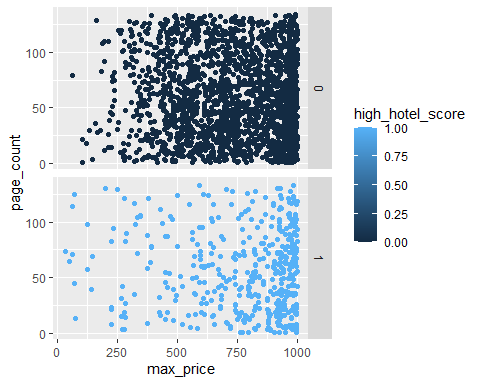
# Generamos el scatter para compromar la homocedasticidad de page\_count cuando dividimos por free\_wifi.  
ggplot(booking, aes(x = max\_price, y = page\_count, color = free\_wifi)) +   
 geom\_point()+  
 facet\_grid(free\_wifi~.)



# Ajustamos el test de fligner-killeen para page\_count y free\_wifi  
fligner.test(page\_count ~ free\_wifi, data = booking)

##   
## Fligner-Killeen test of homogeneity of variances  
##   
## data: page\_count by free\_wifi  
## Fligner-Killeen:med chi-squared = 0.40063, df = 1, p-value = 0.5268

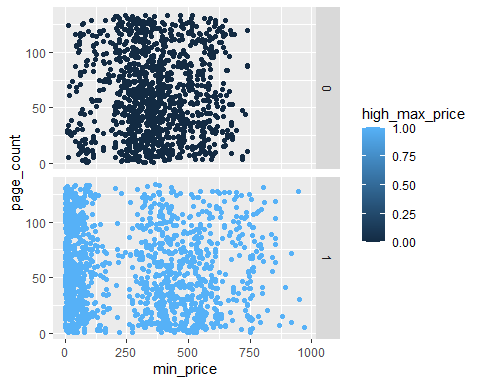
# Generamos la variable para el contraste que realizaremos en la siguiente sección. Si el hotel score es mayor a 9 se marca con una flag  
booking <- booking %>%  
 mutate(high\_hotel\_score = ifelse(hotel\_score >= 9, 1, 0))  
  
# Ploteamos el scatter para comprobar la homocedasticidad de page\_count cuando dividimos por la nueva variable.  
ggplot(booking, aes(x = max\_price, y = page\_count, color = high\_hotel\_score)) +   
 geom\_point()+  
 facet\_grid(high\_hotel\_score~.)



# Ajustamos el test de fligner-killeen para page\_count y la nueva variable  
fligner.test(page\_count ~ high\_hotel\_score, data = booking)

##   
## Fligner-Killeen test of homogeneity of variances  
##   
## data: page\_count by high\_hotel\_score  
## Fligner-Killeen:med chi-squared = 2.8946, df = 1, p-value = 0.08888

# Generamos la variable para el contraste que realizaremos en la siguiente sección. Si el precio máximo supera la media de precios maximos  
# pone un flags  
booking <- booking %>%  
 mutate(high\_max\_price = ifelse(max\_price >= mean(max\_price), 1, 0))  
  
# Ploteamos el scatter para comprobar la homocedasticidad de page\_count cuando dividimos por la nueva variable.  
ggplot(booking, aes(x = min\_price, y = page\_count, color = high\_max\_price)) +   
 geom\_point()+  
 facet\_grid(high\_max\_price~.)



# Ajustamos el test de fligner-killeen para page\_count y la nueva variable  
fligner.test(page\_count ~ high\_max\_price, data = booking)

##   
## Fligner-Killeen test of homogeneity of variances  
##   
## data: page\_count by high\_max\_price  
## Fligner-Killeen:med chi-squared = 0.22846, df = 1, p-value = 0.6327

Tras realizar tests a un gran número de variables de interés (principalmente las que van a ser sujeto de estudio para los tests estadísticos, es decir, air\_conditioning, free\_wifi y dos variables nuevas que hemos creado, high\_hotel\_score y high\_max\_price), se ha determinado que existe homogeneidad de varianzas, por lo que no nos encontraremos con problemas de heterogeneidad a la hora de realizar pruebas de hipótesis.

#### 4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos

En este apartado se aplicarán distintas pruebas estadísticas para comparar los distintos grupos de datos de los que disponemos y ver cual es su relación con la variable objetivo page\_count. Esto nos permitirá evaluar que variables tienen un impacto significativo sobre la variable objetivo y cuales no. A continuación, se recogen unas pocas hipótesis, aunque sería posible evaluar otras que también pareciesen interesantes, dado el carácter académico de este trabajo nos centraremos exclusivamente en estudiar la influencia de las variables:

* **air\_conditioning**
* **has\_free\_cancelation**
* **hotel\_score**
* **max\_price**

Las hipótesis que trataremos de verificar serán las siguientes:

* Aquellos hoteles con aire acondicionado tendrán una mejor posición (page\_count inferior) a los que no dispongan de este.
* Los hoteles con cancelación gratuita tendrán una posición diferente a los que no dispongan de esta.
* Los hoteles con una puntuación global superior a un 9 tendrán un mejor posicionamiento.
* Los hoteles con un precio máximo superior a la media tendrán un peor posicionamiento.

A continuación se estudiará cada una de estas hipótesis y el resultado obtenido.

**Hipótesis 1** La primera hipótesis evalúa la influencia de que un hotel disponga de aire acondicionado en el posicionamiento del hotel, suponiendo que un hotel con aire acondicionado tendrá un posicionamiento mejor, es decir menor **page\_count** que un hotel que no disponga de aire acondicionado. A continuación se plantea la hipótesis nula () y la alternativa ().

* La posición global de un hotel con aire acondicionado y uno sin serán iguales.
* La posición global de un hotel con aire acondicionado será mejor (inferior) a la de uno sin.

O en otras palabras:

En este caso se trata de un **contraste de hipótesis de dos muestras independientes sobre la media con varianzas no conocidas**. Aunque previamente no se ha cumplido la normalidad de los datos por el Teorema del Límite Central podemos considerar la normalidad de los datos y aplicar un test paramétrico. En este caso aplicaremos un **test unilateral por la izquierda**, suponiendo (por lo que se ha visto en el apartado anterior) que las varianzas son iguales, por lo que el estadístico de contraste será:

# Buscamos el conjunto de datos de estudio:  
X1 <- booking %>%  
 filter(air\_conditioning==1) %>%  
 dplyr::select("page\_count")  
X1 <- as.vector(unlist(X1),'numeric')  
  
X2 <- booking %>%  
 filter(air\_conditioning==0) %>%  
 dplyr::select("page\_count")  
X2 <- as.vector(unlist(X2),'numeric')  
  
# Buscamos la varianza de ambas variables  
var1 <- var(X1)  
var2 <- var(X2)  
  
# Definimos el nivel de confianza al 95%  
alfa <- 0.05  
  
# Buscamos las medias y la el número de registros en cada variable  
mean1 <- mean(X1); n1 <- length(X1)  
mean2 <- mean(X2); n2 <- length(X2)  
  
# Calculamos el valor observado y el valor crítico  
tobs <- (mean1-mean2)/sqrt((1/n1 + 1/n2) \* ((n1-1)\*var1 + (n2-1)\*var2)/(n1+n2-2))  
tcrit <- pt(1-alfa, df = (n1+n2-2))  
  
# Finalmente calculamos el pvalue  
pvalue <- pt(tobs, lower.tail=TRUE, df = (n1+n2-2))  
print(paste("The p-value obtained is: ", pvalue))

## [1] "The p-value obtained is: 5.33600165433049e-13"

# Mostramos las medias de las dos variables  
print(paste("The mean of the group 1 is: ", mean(X1)))

## [1] "The mean of the group 1 is: 55.4309293146753"

print(paste("The mean of the group 2 is: ", mean(X2)))

## [1] "The mean of the group 2 is: 64.871609403255"

Se obtiene un p-value muy cercano a 0, y muy inferior a 0.05, por lo que podemos rechazar la hipótesis nula y considerar que **la presencia de aire acondicionado en un hotel mejorará el posicionamiento de este en el buscador de Booking**. Esto se observa también al obtener una media de posicionamiento inferior en más de 9 posiciones en aquellos hoteles con aire acondicionado.

**Hipótesis 2** Esta segunda hipótesis trata de evaluar la influencia de la cancelación gratuita, la cual para aportar una mayor diversidad en los análisis realizados, puede estar asociada a una posición distinta (inferior o superior) que la de un hotel sin cancelación gratuita. Definimos esta hipótesis mediante las hipótesis nula () y alternativa () siguientes:

* La posición global de un hotel con cancelación gratuita y uno sin serán iguales.
* La posición global de un hotel con cancelación gratuita será distinta a la de uno sin.

O en otras palabras:

En este caso se trata de un **contraste de hipótesis de dos muestras independientes sobre la media con varianzas no conocidas**, al igual que en el caso anterior por el TLC podemos aceptar la normalidad de los datos y aplicar un test paramétrico, en concreto un **test bilateral**.

# Buscamos el conjunto de datos de estudio:  
X1 <- booking %>%  
 filter(has\_free\_cancelation==1) %>%  
 dplyr::select("page\_count")  
X1 <- as.vector(unlist(X1),'numeric')  
  
X2 <- booking %>%  
 filter(has\_free\_cancelation==0) %>%  
 dplyr::select("page\_count")  
X2 <- as.vector(unlist(X2),'numeric')  
  
# Buscamos la varianza de ambas variables  
var1 <- var(X1)  
var2 <- var(X2)  
  
# Definimos el nivel de confianza al 95%  
alfa <- 0.05  
  
# Buscamos las medias y la el número de registros en cada variable  
mean1 <- mean(X1); n1 <- length(X1)  
mean2 <- mean(X2); n2 <- length(X2)  
  
# Calculamos el valor observado y el valor crítico  
tobs <- (mean1-mean2)/sqrt((1/n1 + 1/n2) \* ((n1-1)\*var1 + (n2-1)\*var2)/(n1+n2-2))  
tcrit <- pt(1-alfa/2, df = (n1+n2-2))  
  
# Finalmente calculamos el pvalue  
pvalue <- pt(tobs, lower.tail=TRUE, df = (n1+n2-2))  
print(paste("The p-value obtained is: ", pvalue))

## [1] "The p-value obtained is: 0.000376181455513205"

# Mostramos las medias de las dos variables  
print(paste("The mean of the group 1 is: ", mean(X1)))

## [1] "The mean of the group 1 is: 57.6696284329564"

print(paste("The mean of the group 2 is: ", mean(X2)))

## [1] "The mean of the group 2 is: 67.3128491620112"

Observamos que se obtiene un p-value inferior a 0.05 por lo que debemos rechazar la hipótesis nula y, por lo tanto, obtenemos que **un hotel con cancelación gratuita obtendrá un posicionamiento distinto al de un hotel sin cancelación gratuita**. En concreto se ha observado que los hoteles con cancelación gratuita obtienen un mejor posicionamiento que aquellos sin.

**Hipótesis 3** A continuación evaluamos la influencia de la puntuación global del hotel. En este caso evaluaremos si un hotel con una puntuación global (**hotel\_score**) superior o igual a un 9.0 obtiene un mejor posicionamiento que un hotel con una peor puntuación. A continuación se plantea la hipótesis nula () y la alternativa ()

Los hoteles con una puntuación global superior a un 9 tendrán un mejor posicionamiento.

* La posición global de un hotel con una puntuación global mayor o igual a 9.0 y uno con una puntuación inferior serán iguales.
* La posición global de un hotel con una puntuación global mayor o igual a 9.0 será mejor (menor) que uno con una puntuación inferior.

O en otras palabras:

En este caso se trata de un **contraste de hipótesis de dos muestras independientes sobre la media con varianzas no conocidas**. Al igual que en los casos anteriores podemos asumir la normalidad de los datos por el TLC. En este caso se aplicará un **test unilateral por la izquierda**.

# Buscamos el conjunto de datos de estudio:  
X1 <- booking %>%  
 filter(high\_hotel\_score == 1) %>%  
 dplyr::select("page\_count")  
X1 <- as.vector(unlist(X1),'numeric')  
  
X2 <- booking %>%  
 filter(high\_hotel\_score == 0) %>%  
 dplyr::select("page\_count")  
X2 <- as.vector(unlist(X2),'numeric')  
  
# Buscamos la varianza de ambas variables  
var1 <- var(X1)  
var2 <- var(X2)  
  
# Definimos el nivel de confianza al 95%  
alfa <- 0.05  
  
# Buscamos las medias y la el número de registros en cada variable  
mean1 <- mean(X1); n1 <- length(X1)  
mean2 <- mean(X2); n2 <- length(X2)  
  
# Calculamos el valor observado y el valor crítico  
tobs <- (mean1-mean2)/sqrt((1/n1 + 1/n2) \* ((n1-1)\*var1 + (n2-1)\*var2)/(n1+n2-2))  
tcrit <- pt(1-alfa, df = (n1+n2-2))  
  
# Finalmente calculamos el pvalue  
pvalue <- pt(tobs, lower.tail=TRUE, df = (n1+n2-2))  
print(paste("The p-value obtained is: ", pvalue))

## [1] "The p-value obtained is: 0.0483871639663221"

# Mostramos las medias de las dos variables  
print(paste("The mean of the group 1 is: ", mean(X1)))

## [1] "The mean of the group 1 is: 55.9346210995542"

print(paste("The mean of the group 2 is: ", mean(X2)))

## [1] "The mean of the group 2 is: 58.5683229813665"

Se obtiene un valor del p-value de 0.048, el cual aún siendo muy cercano a 0.05 que utilizamos para el intérvalo de confianza es inferior, por lo que se debe rechazar la hipótesis nula y podemos decir que **un hotel con una puntuación superior a 9.0 obtendrá un mejor posicionamiento que el resto de hoteles**. No obstante, por el valor obtenido podemos determinar que esta variable no es tan influyente como el aire acondicionado o la cancelación gratuita.

**Hipótesis 4** Finalmente evaluamos la influencia del precio máximo del hotel considerando que un hotel con un mayor precio máximo tendrá un peor posicionamiento. Para ello consideraremos las siguientes hipótesis nula () y alternativa ()

Los hoteles con un precio máximo superior a la media tendrán un peor posicionamiento.

* La posición global de un hotel con precio máximo superior a la media y uno inferior serán iguales.
* La posición global de un hotel con precio máximo superior a la media será peor (superior) a la de uno con precio inferior

O en otras palabras:

En este caso también realizaremos un **contraste de hipótesis de dos muestras independientes sobre la media con varianzas no conocidas**. Como se ha visto en los casos anteriores, por el TLC podemos asumir la normalidad de los datos y aplicaremos un test paramétrico, en este caso emplearemos un **test unilateral por la derecha**.

# Buscamos el conjunto de datos de estudio:  
X1 <- booking %>%  
 filter(high\_max\_price == 1) %>%  
 dplyr::select("page\_count")  
X1 <- as.vector(unlist(X1),'numeric')  
  
X2 <- booking %>%  
 filter(high\_max\_price == 0) %>%  
 dplyr::select("page\_count")  
X2 <- as.vector(unlist(X2),'numeric')  
  
# Buscamos la varianza de ambas variables  
var1 <- var(X1)  
var2 <- var(X2)  
  
# Definimos el nivel de confianza al 95%  
alfa <- 0.05  
  
# Buscamos las medias y la el número de registros en cada variable  
mean1 <- mean(X1); n1 <- length(X1)  
mean2 <- mean(X2); n2 <- length(X2)  
  
# Calculamos el valor observado y el valor crítico  
tobs <- (mean1-mean2)/sqrt((1/n1 + 1/n2) \* ((n1-1)\*var1 + (n2-1)\*var2)/(n1+n2-2))  
tcrit <- pt(1-alfa, df = (n1+n2-2))  
  
# Finalmente calculamos el pvalue  
pvalue <- pt(tobs, lower.tail=FALSE, df = (n1+n2-2))  
print(paste("The p-value obtained is: ", pvalue))

## [1] "The p-value obtained is: 0.979867959556137"

# Mostramos las medias de las dos variables  
print(paste("The mean of the group 1 is: ", mean(X1)))

## [1] "The mean of the group 1 is: 56.9824644549763"

print(paste("The mean of the group 2 is: ", mean(X2)))

## [1] "The mean of the group 2 is: 59.4509254066181"

En este caso se ha obtenido un p-value de 0.98, por lo que se acepta la hipótesis nula y obtenemos que **los hoteles con un precio máximo superior a la media no obtiene un peor posicionamiento que el resto de hoteles**. Es más en este caso se ha observado que que la media de posicionamiento de los hoteles con un precio máximo superior a la media ha sido ligeramente mejor al resto.

#### 4.4 Aplicación de funciones lineales para resolución del problema

Una vez analizados los distintos contrastes de hipótesis y hayada la correlación entre las distintas variables, tenemos un conocimiento del conjunto de datos lo suficientemente extenso como para analizar mediante una función lineal el posicionamiento de un hotel en el buscador de Booking.

Tras los resultados que se han obtenido observamos que hay distintas variables que pueden tener una mayor influencia sobre nuestra variable objetivo. Por un lado, si nos basamos en lo obtenido tras el análisis de correlación hayamos que variables como **min\_price**, **comfort\_score**, **cleanliness\_score**, **safe** o **hotel\_score** entre otros tienen una gran relevancia. Por otro lado, variables como **air\_conditioning** o **has\_free\_cancelation** han mostrado una relevancia significativa durante el contraste de hipótesis.

En este apartado comenzaremos tratando de ajustar una regresión lineal simple aproximando la variable dependiente **page\_count** con la variable explicativa que ha proporcionado una mayor correlación, que ha resultado ser **comfort\_score**.

#Create the linear model  
mrl <- lm(page\_count~comfort\_score, data=booking)  
  
#Evaluate the data  
summary(mrl)

##   
## Call:  
## lm(formula = page\_count ~ comfort\_score, data = booking)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -59.376 -32.415 -4.118 30.495 79.714   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 95.014 8.523 11.149 < 2e-16 \*\*\*  
## comfort\_score -4.258 0.981 -4.341 1.46e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 37.33 on 3891 degrees of freedom  
## Multiple R-squared: 0.004819, Adjusted R-squared: 0.004563   
## F-statistic: 18.84 on 1 and 3891 DF, p-value: 1.458e-05

Obtenemos un valor de R-squared de 0.005, lo que es un valor muy bajo y representa una aproximación muy mala de la variable dependiente.

Con el objetivo de mejorar los resultados obtenidos se realizará una regresión lineal múltiple que considere todas las puntuaciones del hotel para predecir la posición de este. Para ello empleamos las variables: **hotel\_score**, **staff\_score**, **facilities\_score**, **cleanliness\_score**, **comfort\_score**, **value\_for\_money\_score**, **location\_score** y **free\_wifi\_score**, para ver como la opinión de los clientes afecta en el posicionamiento de este en Booking.

#Create the linear model  
mrl\_scores <- lm(page\_count~hotel\_score + staff\_score + facilities\_score + cleanliness\_score + comfort\_score + value\_for\_money\_score + location\_score + free\_wifi\_score, data=booking)  
  
#Evaluate the data  
summary(mrl\_scores)

##   
## Call:  
## lm(formula = page\_count ~ hotel\_score + staff\_score + facilities\_score +   
## cleanliness\_score + comfort\_score + value\_for\_money\_score +   
## location\_score + free\_wifi\_score, data = booking)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -66.867 -32.255 -4.229 30.134 84.064   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 98.8740 11.8693 8.330 < 2e-16 \*\*\*  
## hotel\_score -5.7102 4.4733 -1.277 0.201851   
## staff\_score 8.8844 1.5676 5.668 1.55e-08 \*\*\*  
## facilities\_score 3.0648 3.2611 0.940 0.347381   
## cleanliness\_score 0.2208 2.7792 0.079 0.936684   
## comfort\_score -2.4574 3.2486 -0.756 0.449424   
## value\_for\_money\_score -8.5827 2.2135 -3.877 0.000107 \*\*\*  
## location\_score -3.2152 1.2112 -2.655 0.007975 \*\*   
## free\_wifi\_score 2.7024 0.7750 3.487 0.000494 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 37.05 on 3884 degrees of freedom  
## Multiple R-squared: 0.02143, Adjusted R-squared: 0.01942   
## F-statistic: 10.63 on 8 and 3884 DF, p-value: 6.894e-15

Se observa que con la regresión lineal multiple R-squared aumenta hasta 0.019, algo mejor que empleando solo una variable, pero aún así los resultados obtenidos son muy malos y no aproximan de una forma correcta la variable dependiente.

De este primer análisis observamos que de todas las variables empleadas, las que han tenido una mayor influencia sobre la variable dependiente han sido **staff\_score**, **value\_for\_money\_score**, **location\_store** y **free\_wifi\_score**. A continuación generamos una regresión lineal múltiple empleando tan solo estas variables.

#Create the linear model  
mrl\_scores <- lm(page\_count~ staff\_score + value\_for\_money\_score + location\_score + free\_wifi\_score, data=booking)  
  
#Evaluate the data  
summary(mrl\_scores)

##   
## Call:  
## lm(formula = page\_count ~ staff\_score + value\_for\_money\_score +   
## location\_score + free\_wifi\_score, data = booking)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -70.946 -32.411 -4.056 29.734 84.276   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 99.4461 11.3863 8.734 < 2e-16 \*\*\*  
## staff\_score 7.3609 1.3427 5.482 4.47e-08 \*\*\*  
## value\_for\_money\_score -11.2927 1.5391 -7.337 2.64e-13 \*\*\*  
## location\_score -3.7267 1.1534 -3.231 0.001244 \*\*   
## free\_wifi\_score 2.5432 0.7651 3.324 0.000895 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 37.06 on 3888 degrees of freedom  
## Multiple R-squared: 0.01995, Adjusted R-squared: 0.01894   
## F-statistic: 19.78 on 4 and 3888 DF, p-value: 3.882e-16

Observamos que habiendo empleado tan solo 4 variables, de las 8 empleadas previamente, el R-squared tan solo se ha visto reducido en 0.0015.

A continuación, mantendremos estas variables que han mostrado los mejores resultados y añadiremos las que habíamos hallado que tenían una mayor correlación con la variable dependiente.

#Create the linear model  
mrl\_complete <- lm(page\_count~staff\_score + value\_for\_money\_score + location\_score + free\_wifi\_score + min\_price + balcony + pet\_friendly + air\_conditioning + private\_bathroom + free\_wifi + swimming\_pool + non\_smoking\_rooms + mean\_flights + length\_description + safe, data=booking)  
  
#Evaluate the data  
summary(mrl\_complete)

##   
## Call:  
## lm(formula = page\_count ~ staff\_score + value\_for\_money\_score +   
## location\_score + free\_wifi\_score + min\_price + balcony +   
## pet\_friendly + air\_conditioning + private\_bathroom + free\_wifi +   
## swimming\_pool + non\_smoking\_rooms + mean\_flights + length\_description +   
## safe, data = booking)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -69.419 -30.664 -4.456 29.302 87.186   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.776e+01 1.193e+01 8.195 3.38e-16 \*\*\*  
## staff\_score 5.482e+00 1.410e+00 3.887 0.000103 \*\*\*  
## value\_for\_money\_score -1.109e+01 1.577e+00 -7.030 2.44e-12 \*\*\*  
## location\_score -2.239e+00 1.214e+00 -1.844 0.065226 .   
## free\_wifi\_score 3.308e+00 7.720e-01 4.285 1.87e-05 \*\*\*  
## min\_price -3.067e-03 2.908e-03 -1.054 0.291739   
## balcony 6.244e+00 1.640e+00 3.807 0.000143 \*\*\*  
## pet\_friendly 1.130e+01 1.640e+00 6.889 6.53e-12 \*\*\*  
## air\_conditioning -1.176e+01 2.391e+00 -4.919 9.06e-07 \*\*\*  
## private\_bathroom -2.128e+00 1.533e+00 -1.388 0.165176   
## free\_wifi 2.832e+00 2.730e+00 1.037 0.299633   
## swimming\_pool -4.447e+00 1.918e+00 -2.318 0.020492 \*   
## non\_smoking\_rooms -1.394e+00 1.671e+00 -0.834 0.404397   
## mean\_flights -2.032e-06 5.142e-07 -3.951 7.91e-05 \*\*\*  
## length\_description 4.923e-02 2.093e-02 2.352 0.018732 \*   
## safe 9.343e-01 1.569e+00 0.596 0.551489   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 36.41 on 3877 degrees of freedom  
## Multiple R-squared: 0.05666, Adjusted R-squared: 0.05301   
## F-statistic: 15.52 on 15 and 3877 DF, p-value: < 2.2e-16

Con todas las variables la R-squared ha aumentado hasta 0.057, que sigue siendo un resultado muy bajo, pero dificil de mejorar con unos datos tan poco correlacionados con la variable objetivo.

De todas las variables empleadas se observa que las que tienen una mayor influencia son: staff\_score, free\_wifi\_score, balcony, pet\_friendly y length\_description que influyen negativamente en el posicionamiento (aumentan las posiciones en las que aparece, es decir más tarde) y por otro lado, air\_conditioning, swimming\_pool, mean\_flights y value\_for\_money\_score, las cuales influyen positivamente (se aparece en posiciones más tempranas).

Finalmente, se aproxima la variable dependiente basándonos exclusivamente en estas variables.

#Create the linear model  
mrl\_final <- lm(page\_count~staff\_score + value\_for\_money\_score + free\_wifi\_score + balcony + pet\_friendly + air\_conditioning + swimming\_pool + mean\_flights + length\_description, data=booking)  
  
#Evaluate the data  
summary(mrl\_final)

##   
## Call:  
## lm(formula = page\_count ~ staff\_score + value\_for\_money\_score +   
## free\_wifi\_score + balcony + pet\_friendly + air\_conditioning +   
## swimming\_pool + mean\_flights + length\_description, data = booking)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -70.224 -30.654 -4.833 29.421 87.936   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.407e+01 1.021e+01 8.236 2.40e-16 \*\*\*  
## staff\_score 5.342e+00 1.351e+00 3.953 7.85e-05 \*\*\*  
## value\_for\_money\_score -1.159e+01 1.527e+00 -7.588 4.05e-14 \*\*\*  
## free\_wifi\_score 3.085e+00 7.574e-01 4.073 4.73e-05 \*\*\*  
## balcony 6.608e+00 1.515e+00 4.361 1.33e-05 \*\*\*  
## pet\_friendly 1.171e+01 1.609e+00 7.276 4.14e-13 \*\*\*  
## air\_conditioning -1.166e+01 1.399e+00 -8.332 < 2e-16 \*\*\*  
## swimming\_pool -3.584e+00 1.867e+00 -1.919 0.0550 .   
## mean\_flights -2.381e-06 4.876e-07 -4.884 1.08e-06 \*\*\*  
## length\_description 5.152e-02 2.046e-02 2.518 0.0118 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 36.42 on 3883 degrees of freedom  
## Multiple R-squared: 0.05461, Adjusted R-squared: 0.05242   
## F-statistic: 24.92 on 9 and 3883 DF, p-value: < 2.2e-16

De esta forma hemos conseguido mantener una R-squared similar a la obtenida previamente, pero empleando exclusivamente 11 variables en lugar de las más de 40 variables iniciales.

Concluyendo esta última sección, concentramos nuestro análisis en torno a max\_price. Anteriormente se ha observado una correlación entre el precio máximo de la habitación y sus características. Por tanto, seguidamente se tratará de predecir dicho precio mediante una regresión lineal múltiple empleando el resto de variables para ello:

#Create the linear model  
mrl\_price <- lm(max\_price~staff\_score + value\_for\_money\_score + location\_score + free\_wifi\_score + balcony + pet\_friendly + air\_conditioning + private\_bathroom + free\_wifi + swimming\_pool + non\_smoking\_rooms + mean\_flights + length\_description + safe, data=booking)  
  
#Evaluate the data  
summary(mrl\_price)

##   
## Call:  
## lm(formula = max\_price ~ staff\_score + value\_for\_money\_score +   
## location\_score + free\_wifi\_score + balcony + pet\_friendly +   
## air\_conditioning + private\_bathroom + free\_wifi + swimming\_pool +   
## non\_smoking\_rooms + mean\_flights + length\_description + safe,   
## data = booking)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -721.58 -126.28 23.68 135.53 449.68   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.330e+02 6.103e+01 -3.817 0.000137 \*\*\*  
## staff\_score 2.545e+01 7.332e+00 3.472 0.000523 \*\*\*  
## value\_for\_money\_score -1.415e+01 8.180e+00 -1.730 0.083669 .   
## location\_score 6.438e+01 6.313e+00 10.199 < 2e-16 \*\*\*  
## free\_wifi\_score -3.767e+00 4.005e+00 -0.941 0.347017   
## balcony -4.120e+01 8.505e+00 -4.844 1.32e-06 \*\*\*  
## pet\_friendly -1.462e+01 8.522e+00 -1.716 0.086235 .   
## air\_conditioning -2.544e+01 1.243e+01 -2.047 0.040683 \*   
## private\_bathroom 7.441e+00 7.973e+00 0.933 0.350736   
## free\_wifi 6.494e+01 1.413e+01 4.597 4.41e-06 \*\*\*  
## swimming\_pool 5.228e+01 9.860e+00 5.303 1.20e-07 \*\*\*  
## non\_smoking\_rooms 3.796e+01 8.675e+00 4.375 1.24e-05 \*\*\*  
## mean\_flights 3.047e-05 2.674e-06 11.396 < 2e-16 \*\*\*  
## length\_description 1.180e+00 1.086e-01 10.867 < 2e-16 \*\*\*  
## safe 5.257e+01 8.101e+00 6.489 9.75e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 189.4 on 3878 degrees of freedom  
## Multiple R-squared: 0.2125, Adjusted R-squared: 0.2097   
## F-statistic: 74.75 on 14 and 3878 DF, p-value: < 2.2e-16

Con ello obtenemos una R-squared de 0.2125, con lo que podemos observar que variables que tienen una mayor correlación con la variable objetivo nos proporcionarán unos resultados más apropiados. A partir de los resultados mostrados podemos mejorar la regresión empleando tan solo aquellas variables que han aportado más información en la regresión como han sido, por un lado **staff\_score**, **location\_score**, **free\_wifi**, **swimming\_pool**, **non\_smoking\_rooms**, **mean\_flights**, **length\_description** y **safe**, que influyen positivamente en el precio máximo y, por otro lado y aunque parezca sorprendente, **balcony** y **air\_conditioning**, que influyen negativamente en el mismo.

A continuación se emplea una regresión lineal múltiple tan solo con estas variables.

#Create the linear model  
mrl\_price <- lm(max\_price~staff\_score + location\_score + balcony + air\_conditioning + free\_wifi + swimming\_pool + non\_smoking\_rooms + mean\_flights + length\_description + safe, data=booking)  
  
#Evaluate the data  
summary(mrl\_price)

##   
## Call:  
## lm(formula = max\_price ~ staff\_score + location\_score + balcony +   
## air\_conditioning + free\_wifi + swimming\_pool + non\_smoking\_rooms +   
## mean\_flights + length\_description + safe, data = booking)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -719.57 -124.30 24.15 135.57 447.64   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.626e+02 5.926e+01 -4.431 9.63e-06 \*\*\*  
## staff\_score 1.466e+01 5.700e+00 2.573 0.0101 \*   
## location\_score 6.222e+01 6.080e+00 10.233 < 2e-16 \*\*\*  
## balcony -4.256e+01 8.418e+00 -5.056 4.48e-07 \*\*\*  
## air\_conditioning -2.479e+01 1.217e+01 -2.038 0.0416 \*   
## free\_wifi 6.266e+01 1.360e+01 4.608 4.20e-06 \*\*\*  
## swimming\_pool 5.252e+01 9.650e+00 5.442 5.58e-08 \*\*\*  
## non\_smoking\_rooms 3.946e+01 8.589e+00 4.594 4.48e-06 \*\*\*  
## mean\_flights 3.071e-05 2.670e-06 11.504 < 2e-16 \*\*\*  
## length\_description 1.153e+00 1.074e-01 10.738 < 2e-16 \*\*\*  
## safe 5.435e+01 7.880e+00 6.896 6.20e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 189.5 on 3882 degrees of freedom  
## Multiple R-squared: 0.2109, Adjusted R-squared: 0.2088   
## F-statistic: 103.7 on 10 and 3882 DF, p-value: < 2.2e-16

Se observa que aun habiendo reducido el número de variables, mantenemos una R-squares ligeramente superior 0.21. Hay que tener en cuenta que los resultados siguen sin ser satisfactorios, ya que una R-squared de 0.21 es bastante baja, no obstante es mucho mejor que lo obtenido tratando de predecir la variable page\_count. Esta mala predicción es producida como se había comentado previamente debido a la baja correlación de las variables con la variable objetivo.

## 5. REPRESENTACIÓN DE LOS RESULTADOS

A lo largo del documento se han presentado representaciones gráficas de los análisis realizados y de los resultados obtenidos.

## 6. RESOLUCIÓN DEL PROBLEMA

Como se ha visto a lo largo del presente documento, se ha procedido a la limpieza del dataset obtenido a través del scrapeo de Booking.com, con la finalidad de intentar determinar qué variables influyen más en el posicionamiento y la determinación del precio de un hotel en dicha web.

En primer lugar, se ha llevado a cabo un profundo proceso de preprocesado de datos en el que se han creado un número significativo de variables, se han eliminado registros poco útiles o con muchos elementos vacíos o se han imputado valores en los casos oportunos y, finalmente, se ha analizado la existencia de outliers, los cuales no han sido eliminados dado que no se han considerado erróneos, pues los valores entraban dentro de la lógica de cada una de las variables.

Tras llevar a cabo distintas pruebas estadísticas, entre las que destacan el análisis de correlación y los contrates de hipótesis (incluyendo de normalidad y de homocedasticidad), y la culminación de la práctica con la aplicación de modelos de regresión lineal. Se obtienen distintas conclusiones:

Inicialmente, del análisis de correlación se ha observado que existen distintas variables correlacionadas entre ellas como, serían los servicios del hotel o las puntuaciones de las distintas categorías, pero ninguna de ellas parecía tener una correlación suficientemente fuerte con la variable objetivo.

Mediante el contraste de hipótesis se ha obtenido que determinadas variables como la presencia de aire acondicionado o la cancelación gratuita de la reserva sí que tenían un impacto sobre el posicionamiento de un hotel en el buscador, concretamente tener estos servicios implicaba, de media, un mejor posicionamiento en el buscador.

Finalmente, se ha tratado de obtener una regresión lineal múltiple que, a partir de distintas variables independientes, nos permitiese predecir la variable dependiente que indica la posición en la página del buscador. Este proceso ha sido complejo y ha resultado en unos resultados muy poco precisos, que aun habiendo conseguido mejorarse afinando el proceso, no servirían para una aproximación real. Esto es debido a una falta de correlación entre las variables dependientes y la objetivo, lo cual nos indica que hacer uso de modelos de regresión lineal limita la precisión de las predicciones que se puede alcanzar. No obstante, puesto que nuestra tarea era una aproximación inferencial, no predictiva, podemos considerar que nuestro proyecto resuelve, aunque de forma limitada, el problema planteado, pues de este proceso se ha obtenido que de las más de 40 variables iniciales que teníamos, con tan solo 9 de estas podemos desarrollar unas breves recomendaciones con respecto a cómo posicionar un hotel en Booking.

Como conclusión, tras el análisis realizado se ha obtenido que las variables que resultan ser más relevantes, influenciando de forma positiva o negativa en el posicionamiento de hotel son: Staff\_Score, Value\_For\_Money\_Score, Free\_Wifi\_Score, Balcony, Pet\_Friendly, Air\_Conditioning, Swimming\_Pool, Mean\_Flights, Length\_Description.

Por otro lado, en cuanto al problema de pricing, hemos sido capaces de determinar que las variables que resultan más relevantes en la determinación del precio son: Staff\_Score, Location\_Score, Free\_Wifi, Swimming\_Pool, Non\_Smoking\_Rooms, Mean\_Flights, Length\_Description, Safe, Balcony y Air\_conditioning. No obstante, nuestros resultados en cuanto a poder predictivo tienen un gran margen de mejoría.

Como sugerencias tras los resultados del análisis podemos decir que, dado un hotel con unas características determinadas, nuestra recomendación para mejorar su posicionamiento pasaría primero por mejorar la descripción del hotel haciéndola más corta y precisa a la vez que se enfoca en una estrategia de mejora de la calidad percibida con respecto al precio (value for money), ya que son dos medidas “gratuitas” que tendrán un gran impacto. En caso de querer invertir dinero se recomendaría hacer una instalación de una piscina o de aire acondicionado (aunque teniendo en cuenta su influencia en el precio se debería tener en consideración las prioridades del hotel).

Si, por otro lado, un hotel busca incrementar sus precios sería recomendable, aunque influenciando así negativamente en el posicionamiento, mejorar el trato al cliente por parte de los empleados e incluir otros servicios como Wifi gratuito o Piscina.