Universidad Peruana de Ciencias Aplicadas

Ciencias de la Computación
Administración De La Información (CC50-2201-CC52)
2022-1

Profesor :

Reyes Silva, Patricia Daniela

Trabajo Parcial

Integrantes:

Lopez Gonzalez, Maria Isabel - U201724423
Trujillo Mori , Jeanpier Alexander - U201523565
Contreras Inostroza, Eduardo Junior - U201414103

Mayo, 2022

CASO DE ANÁLISIS

El dataset que vamos a analizar contiene información de reservas de dos hoteles ubicados en Portugal uno es un hotel dentro de la ciudad de Lisboa y el otro un resort en la región de Algarvel. La información de estas reservas son desde el 1 de julio del 2015 hasta el 31 de agosto del año 2017. Entre los datos relevantes en el dataset se encuentran la fecha que se realizó la reserva, duración de la misma, cantidad de personas, niños y/o bebés, entre otros. Los datos se originan del artículo científico Hotel booking demand data sets escrito por Nuno Antonio, Ana Almeida, y Luis Nunes. El cual fue publicado por la revista Data in Brief, Volumen 22 en Febrero del 2019.

CASOS DE USO APLICABLES

- Desde la perspectiva empresarial, podemos tener un record de lo que sucede en la empresa a nivel del comportamiento que una persona tiene al reservar espacios en un hotel, y estos datos nos puede ser útil para crear perfiles de comportamiento, los cuales pueden ser ser usados para predecir el comportamiento de un cliente futuro o la detección de patrones que cumplen los clientes a la hora de hacer reservas, esta estrategia podría significar una ventaja competitiva frente a los Hoteles que no hacen esta toma de datos o esta generación de perfiles.

- Desde la perspectiva del cliente, puede tener la sensación de una atención personalizada gracias a que sus datos se están guardando y perfiles acorde se están creando.

CONJUNTO DE DATOS (DATA SET)

Columna	Descripción	Tipo
hotel	Tipo de hotel donde se hizo la reserva puede ser **Resort Hotel** o **City Hotel**	Categórica
is_canceled	Valor indicando si la reserva fue cancelada (1) o no (0)	Categórica
lead_time	Cantidad de días que transcurrieron entre crear la reserva en el sistema y el día de llegada o cancelación.	Numérico
arrival_date_year	Año de llegada en la reserva	Numérico
arrival_date_month	Mes de llegada en la reserva	Numérico
arrival_date_week_number	Semana del año para la reserva	Numérico
arrival_date_day_of_month	Día de llegada en la reserva	Numérico
stays_in_weekend_nights	Cantidad de noches de fin de semana (sábado y domingo) que el huésped reservo	Numérico
stays_in_week_nights	Cantidad de noches de dia de semana (lunes a viernes) que el huésped reservo	Numérico
adults	Número de adultos	Numérico
children	Número de niños	Numérico
babies	Número de bebés	Numérico
meal	Tipo de comida reservada **SC o no data** (no hay comida), **BB** - (desayuno), **HB** - (usualmente desayuno y otra comida) y **FB** (las tres comidas).	Categórica
country	País de origen, representado en formato ISO 3155–3:2013	Categórica
market_segment	Segmentación de Mercado	Categórica
distribution_channel	Indica el canal distribución u origen de las reservas. Las principales categorías son **Direct** (el huésped directamente), **Corporativo**, **TA** (Agencias de Viaje) o **TO** (Tours)	Categórica
is_repeated_guest	Valor indicando si es un huésped a reservado anteriormente (1) o no (0)	Categórica
previous_cancellations	Número de reservas canceladas por el cliente antes de la reserva actual.	Numérico
previous_bookings_not_canceled	Número de reservas no canceladas por el cliente	Numérico

	antes de la reserva actual..	
reserved_room_type	Código de la habitación reservada.	Categórica
assigned_room_type	Código de la habitación asignada.	Categórica
booking_changes	Número de cambios en la reserva desde el momento que se creó.	Numérico
deposit_type	Indica si el huésped hizo un depósito para garantizar la reserva. Puede tener tres valores: **No Deposit** (no se hizo ningún depósito), **Non Refund** (se hizo un depósito por el costo total) y **Refundable** (Se hizo un pago menor al costo total).	Categórica
agent	ID de la agencia de viaje que hizo la reserva	Categórica
company	ID de la empres que hizo la reserva	Categórica
days_in_waiting_list	Días en lista de espera antes de confirmar la reserva con el cliente	Numérico
customer_type	Tipo de cliente, puede ser uno de cuatro tipos, **Contract** (reserva por un contra), **Group** (reserva asociada a un grupo), **Transient** (reserva no asociada a un grupo) y **Transient-party** (reserva asociada a otra reserva)	Categórica
adr	Average Daily Rate	Numérico
required_car_parking_spaces	Número de puestos de estacionamientos solicitados por el cliente	Numérico
total_of_special_requests	Número de solicitudes especiales por parte del cliente	Numérico
reservation_status	Último status de la reserva	Categórica
reservation_status_date	Fecha de último cambio de estatus de la reserva	Fecha

ANÁLISIS EXPLORATORIO DE DATOS

CARGAR DATO

Cargamos todos los datos en el data set *hotel_booking_missing.csv*. Usamos los parametros `header = TRUE` para indicar que nuestro dataset tiene una columna inicial con la descripttion de los datos y el parametro `stringAsFactors = TRUE` donde todas las columnas con el tipo char son factor.

```
hotel_bookings <- read.csv("hotel_bookings_miss.csv", header=TRUE,
stringsAsFactors = TRUE)

names(hotel_bookings)[names(hotel_bookings) == 'ï..hotel'] <- 'hotel'
```

INSPECCIONAR DATOS

Visualizar los datos en el dataset

```
names(hotel_bookings)
str(hotel_bookings)
```

```
 [1] "hotel"                         "is_canceled"                   "lead_time"
 [4] "arrival_date_year"             "arrival_date_month"            "arrival_date_week_number"
 [7] "arrival_date_day_of_month"     "stays_in_weekend_nights"       "stays_in_week_nights"
[10] "adults"                        "children"                      "babies"
[13] "meal"                          "country"                       "market_segment"
[16] "distribution_channel"          "is_repeated_guest"             "previous_cancellations"
[19] "previous_bookings_not_canceled" "reserved_room_type"           "assigned_room_type"
[22] "booking_changes"               "deposit_type"                  "agent"
[25] "company"                       "days_in_waiting_list"          "customer_type"
[28] "adr"                           "required_car_parking_spaces"   "total_of_special_requests"
[31] "reservation_status"            "reservation_status_date"
'data.frame':   119390 obs. of  32 variables:
 $ hotel                         : Factor w/ 2 levels "City Hotel","Resort Hotel": 2 2 2 2 2 2 2 2 2 2 ...
 $ is_canceled                   : int  0 0 0 0 0 0 0 0 1 1 ...
 $ lead_time                     : int  342 737 7 13 14 14 0 9 85 75 ...
 $ arrival_date_year             : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
 $ arrival_date_month            : Factor w/ 12 levels "April","August",..: 6 6 6 6 6 6 6 6 6 6 ...
 $ arrival_date_week_number      : int  27 27 27 27 27 27 27 27 27 27 ...
 $ arrival_date_day_of_month     : int  1 1 1 1 1 1 1 1 1 1 ...
 $ stays_in_weekend_nights       : int  NA 0 0 0 0 0 0 0 0 0 ...
 $ stays_in_week_nights          : int  0 0 1 1 2 2 2 2 3 3 ...
 $ adults                        : int  2 2 1 1 2 2 2 2 2 2 ...
 $ children                      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ babies                        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ meal                          : Factor w/ 5 levels "BB","FB","HB",..: 1 1 1 1 1 1 1 2 1 3 ...
 $ country                       : Factor w/ 178 levels "ABW","AGO","AIA",..: 137 137 60 60 60 60 137 137 137 137
...
 $ market_segment                : Factor w/ 8 levels "Aviation","Complementary",..: 4 4 4 3 7 7 4 4 7 6 ...
 $ distribution_channel          : Factor w/ 5 levels "Corporate","Direct",..: 2 2 2 1 4 4 2 2 4 4 ...
 $ is_repeated_guest             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ previous_cancellations        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ previous_bookings_not_canceled: int  0 0 0 0 0 0 0 0 0 0 ...
 $ reserved_room_type            : Factor w/ 10 levels "A","B","C","D",..: 3 3 1 1 1 1 3 3 1 4 ...
 $ assigned_room_type            : Factor w/ 12 levels "A","B","C","D",..: 3 3 3 1 1 1 3 3 1 4 ...
 $ booking_changes               : int  3 4 0 0 0 0 0 0 0 0 ...
 $ deposit_type                  : Factor w/ 3 levels "No Deposit","Non Refund",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ agent                         : Factor w/ 334 levels "1","10","103",..: 334 334 334 157 103 103 334 156 103 40
...
 $ company                       : Factor w/ 353 levels "10","100","101",..: 353 353 353 353 353 353 353 353 353
353 ...
 $ days_in_waiting_list          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ customer_type                 : Factor w/ 4 levels "Contract","Group",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ adr                           : num  0 0 75 75 98 ...
 $ required_car_parking_spaces   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ total_of_special_requests     : int  0 0 0 0 1 1 0 1 1 0 ...
 $ reservation_status            : Factor w/ 3 levels "Canceled","Check-Out",..: 2 2 2 2 2 2 2 2 1 1 ...
 $ reservation_status_date       : Factor w/ 926 levels "1/1/2015","1/1/2016",..: 669 669 702 702 735 735 735 735
570 449 ...
```

```
summary(hotel_bookings)
```

```
         hotel        is_canceled       lead_time       arrival_date_year arrival_date_month arrival_date_week_number
 City Hotel  :79330   Min.   :0.0000   Min.   :  0.0   Min.   :2015      August :13877       Min.   : 1.00
 Resort Hotel:40060   1st Qu.:0.0000   1st Qu.: 18.0   1st Qu.:2016      July   :12661       1st Qu.:16.00
                      Median :0.0000   Median : 69.0   Median :2016      May    :11791       Median :28.00
                      Mean   :0.3704   Mean   :104.0   Mean   :2016      October:11160       Mean   :27.16
                      3rd Qu.:1.0000   3rd Qu.:160.0   3rd Qu.:2017      April  :11089       3rd Qu.:38.00
                      Max.   :1.0000   Max.   :737.0   Max.   :2017      June   :10939       Max.   :53.00
                                       NA's   :21      NA's   :6         (Other):47873       NA's   :25
 arrival_date_day_of_month stays_in_weekend_nights stays_in_week_nights     adults          children
 Min.   : 1.0              Min.   : 0.0000        Min.   : 0.0          Min.   : 0.000   Min.   : 0.0000
 1st Qu.: 8.0              1st Qu.: 0.0000        1st Qu.: 1.0          1st Qu.: 2.000   1st Qu.: 0.0000
 Median :16.0              Median : 1.0000        Median : 2.0          Median : 2.000   Median : 0.0000
 Mean   :15.8              Mean   : 0.9275        Mean   : 2.5          Mean   : 1.856   Mean   : 0.1039
 3rd Qu.:23.0              3rd Qu.: 2.0000        3rd Qu.: 3.0          3rd Qu.: 2.000   3rd Qu.: 0.0000
 Max.   :31.0              Max.   :19.0000        Max.   :50.0          Max.   :55.000   Max.   :10.0000
 NA's   :7                 NA's   :25             NA's   :12            NA's   :12       NA's   :4
     babies                meal          country            market_segment  distribution_channel is_repeated_guest
 Min.   : 0.00000   BB        :92310   PRT    :48590   Online TA     :56477   Corporate: 6677      Min.   :0.00000
 1st Qu.: 0.00000   FB        :  798   GBR    :12129   Offline TA/TO:24219   Direct   :14645      1st Qu.:0.00000
 Median : 0.00000   HB        :14463   FRA    :10415   Groups        :19811   GDS      :  193      Median :0.00000
 Mean   : 0.00795   SC        :10650   ESP    : 8568   Direct        :12606   TA/TO    :97870      Mean   :0.03191
 3rd Qu.: 0.00000   Undefined: 1169   DEU    : 7287   Corporate     : 5295   Undefined:    5      3rd Qu.:0.00000
 Max.   :10.00000                     ITA    : 3766   Complementary:  743                         Max.   :1.00000
 NA's   :32                           (Other):28635   (Other)       :  239
 previous_cancellations previous_bookings_not_canceled reserved_room_type assigned_room_type booking_changes
 Min.   : 0.00000       Min.   : 0.0000                A      :85994      A      :74053      Min.   : 0.0000
 1st Qu.: 0.00000       1st Qu.: 0.0000                D      :19201      D      :25322      1st Qu.: 0.0000
 Median : 0.00000       Median : 0.0000                E      : 6535      E      : 7806      Median : 0.0000
 Mean   : 0.08712       Mean   : 0.1371                F      : 2897      F      : 3751      Mean   : 0.2211
 3rd Qu.: 0.00000       3rd Qu.: 0.0000                G      : 2094      G      : 2553      3rd Qu.: 0.0000
 Max.   :26.00000       Max.   :72.0000                B      : 1118      C      : 2375      Max.   :21.0000
                                                       (Other): 1551      (Other): 3530
     deposit_type        agent          company       days_in_waiting_list         customer_type
 No Deposit:104641   9      :31961   NULL   :112593   Min.   :  0.000      Contract        : 4076
 Non Refund: 14587   NULL   :16340   40     :   927   1st Qu.:  0.000      Group           :  577
 Refundable:   162   240    :13922   223    :   784   Median :  0.000      Transient       :89613
                     1      : 7191   67     :   267   Mean   :  2.321      Transient-Party:25124
                     14     : 3640   45     :   250   3rd Qu.:  0.000
                     7      : 3539   153    :   215   Max.   :391.000
                     (Other):42797   (Other):  4354   NA's   :7
      adr          required_car_parking_spaces total_of_special_requests reservation_status reservation_status_date
 Min.   :  -6.38   Min.   :0.00000             Min.   :0.0000            Canceled :43017    10/21/2015:  1461
 1st Qu.:  69.29   1st Qu.:0.00000             1st Qu.:0.0000            Check-Out:75166    7/6/2015  :   805
 Median :  94.58   Median :0.00000             Median :0.0000            No-Show  : 1207    11/25/2016:   790
 Mean   : 101.83   Mean   :0.06252             Mean   :0.5714                               1/1/2015  :   763
 3rd Qu.: 126.00   3rd Qu.:0.00000             3rd Qu.:1.0000                               1/18/2016 :   625
 Max.   :5400.00   Max.   :8.00000             Max.   :5.0000                               7/2/2015  :   469
                                                                                            (Other)   :114477
```

Actualizar los tipos necesarios a factores

```
hb_data <- hotel_bookings
hb_data$is_canceled <- as.factor(hb_data$is_canceled)

hb_data$arrival_date_week_number <- as.factor(hb_data$arrival_date_week_number)
hb_data$arrival_date_year <- as.factor(hb_data$arrival_date_year)
hb_data$arrival_date_day_of_month  <-
as.factor(hb_data$arrival_date_day_of_month)

hb_data$is_repeated_guest <- as.factor(hb_data$is_repeated_guest)

hb_data$reservation_status_date <- as.Date(hb_data$reservation_status_date,
"%m/%d/%Y")

hb_data[sapply(hb_data, is.character)] <-
  lapply(hb_data[sapply(hb_data, is.character)], as.factor)
str(hb_data)
```

```
'data.frame':   119390 obs. of  32 variables:
 $ hotel                         : Factor w/ 2 levels "City Hotel","Resort Hotel": 2 2 2 2 2 2 2 2 2 2 ...
 $ is_canceled                   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 2 2 ...
 $ lead_time                     : int  342 737 7 13 14 14 0 9 85 75 ...
 $ arrival_date_year             : Factor w/ 3 levels "2015","2016",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ arrival_date_month            : Factor w/ 12 levels "April","August",..: 6 6 6 6 6 6 6 6 6 6 ...
 $ arrival_date_week_number      : Factor w/ 53 levels "1","2","3","4",..: 27 27 27 27 27 27 27 27 27 27 ...
 $ arrival_date_day_of_month     : Factor w/ 31 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ stays_in_weekend_nights       : int  NA 0 0 0 0 0 0 0 0 0 ...
 $ stays_in_week_nights          : int  0 0 1 1 2 2 2 2 3 3 ...
 $ adults                        : int  2 2 1 1 2 2 2 2 2 2 ...
 $ children                      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ babies                        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ meal                          : Factor w/ 5 levels "BB","FB","HB",..: 1 1 1 1 1 1 1 2 1 3 ...
 $ country                       : Factor w/ 178 levels "ABW","AGO","AIA",..: 137 137 60 60 60 60 137 137 137 137 ...
 $ market_segment                : Factor w/ 8 levels "Aviation","Complementary",..: 4 4 4 3 7 7 4 4 7 6 ...
 $ distribution_channel          : Factor w/ 5 levels "Corporate","Direct",..: 2 2 2 1 4 4 2 2 4 4 ...
 $ is_repeated_guest             : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ previous_cancellations        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ previous_bookings_not_canceled: int  0 0 0 0 0 0 0 0 0 0 ...
 $ reserved_room_type            : Factor w/ 10 levels "A","B","C","D",..: 3 3 1 1 1 1 3 3 1 4 ...
 $ assigned_room_type            : Factor w/ 12 levels "A","B","C","D",..: 3 3 3 1 1 1 3 3 1 4 ...
 $ booking_changes               : int  3 4 0 0 0 0 0 0 0 0 ...
 $ deposit_type                  : Factor w/ 3 levels "No Deposit","Non Refund",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ agent                         : Factor w/ 334 levels "1","10","103",..: 334 334 334 157 103 103 334 156 103 40 ...
 $ company                       : Factor w/ 353 levels "10","100","101",..: 353 353 353 353 353 353 353 353 353 353 .
 $ days_in_waiting_list          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ customer_type                 : Factor w/ 4 levels "Contract","Group",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ adr                           : num  0 0 75 75 98 ...
 $ required_car_parking_spaces   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ total_of_special_requests     : int  0 0 0 0 1 1 0 1 1 0 ...
 $ reservation_status            : Factor w/ 3 levels "Canceled","Check-Out",..: 2 2 2 2 2 2 2 2 1 1 ...
 $ reservation_status_date       : Date, format: "2015-07-01" "2015-07-01" "2015-07-02" "2015-07-02" ...
```

PRE-PROCESAR DATOS

Identificación de datos faltantes

```
columnns_NA_values <- function(x){
 count = 0

 for(i in 1:ncol(x)) {
   if (colSums(is.na(x[i]))){
      cat("NA values:",colSums(is.na(x[i])), "    \tColumn",colnames(x[i]),"\n")
      count = count + 1
   }
 }

 cat("Columns with NA values: ", count, "\n\n")
}

columnns_wempty_values <- function(x){
 count = 0

 for(i in 1:ncol(x)) {
   if (isTRUE(colSums(x[i]==""))){
      cat("NA values:",colSums(x[i]==""), "\tColumn",colnames(x[i]),"\n")
      count = count + 1
   }
 }

 cat("Columns with empty values: ", count, "\n\n")
}

columnns_NA_values(hb_data)
columnns_wempty_values(hb_data)
```

```
NA values: 21       Column lead_time
NA values: 6        Column arrival_date_year
NA values: 25       Column arrival_date_week_number
NA values: 7        Column arrival_date_day_of_month
NA values: 25       Column stays_in_weekend_nights
NA values: 12       Column stays_in_week_nights
NA values: 12       Column adults
NA values: 4        Column children
NA values: 32       Column babies
NA values: 7        Column days_in_waiting_list
NA values: 72423    Column reservation_status_date
Columns with NA values:  11

Columns with empty values:  0
```

Logramos identificar que existen 10 columnas con valores faltantes y ninguna de las columnas tiene valores vacíos. Debido a que son varios datos que tiene valores faltantes nos enfocaremos en las columnas más relevantes a las preguntas que deseamos resolver.

children y babies

En el caso del número de niños y bebés en las reservas podemos asumir de una forma casi certera que si el valor de la reserva es NA la reserva no cuenta con ninguno de estos en la misma. Serían 36 filas que modificaremos y le daremos el valor de 0.

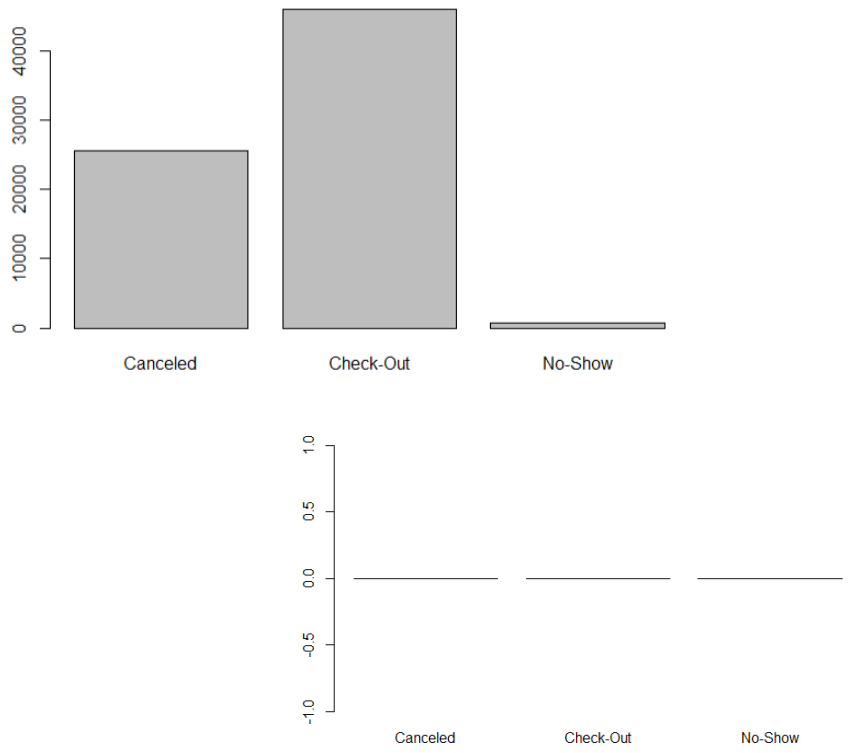
```
hb_data[is.na(hb_data$children ) | is.na(hb_data$babies ),]
hb_data[is.na(hb_data$children ),]$children <- 0
hb_data[is.na(hb_data$babies ),]$babies <- 0
```

reservation_status_date

En el caso de reservation status, la cantidad de datos faltantes es muy significativa para ignorar, para ello primero visualizamos por tipo de reservas que cantidad de datos no tienen una fecha en el status date. Dependiendo del reservation_status lo manejamos diferente. Si el reservation_status es Check-Out el último día será la suma de la fecha de reserva más los días de la estadía indicados en la reserva. Si el status es No-Show, la fecha será la fecha indicada en la reserva. Y por último como las tablas están ordenadas por la fecha de reserva asumimos que el orden de la reserva es cronológico por ello utilizaremos la función fill() para completar el resto de las fechas faltantes para las reservas con estatus Canceled.

```
reserve_status <-
hb_data[is.na(hb_data$reservation_status_date),]$reservation_status
bp <- barplot(table(reserve_status)) + theme_bw()

temp_hb <- hb_data

#Update reservation_status == Check-Out
check_out_bool <- is.na(hb_data$reservation_status_date) &
hb_data$reservation_status == 'Check-Out'
nd_check_out <- hb_data[check_out_bool, ]

reservation_date <- paste(nd_check_out$arrival_date_year,
nd_check_out$arrival_date_month, nd_check_out$arrival_date_day_of_month)
last_update <- as.Date(reservation_date, format = "%Y %B %d") +
nd_check_out$stays_in_weekend_nights + nd_check_out$stays_in_week_nights

temp_hb[check_out_bool, ]$reservation_status_date <- last_update
```

```
#Update reservation_status == No-Show
no_show_bool <- is.na(hb_data$reservation_status_date) &
hb_data$reservation_status == 'No-Show'
nd_no_show <- hb_data[no_show_bool, ]

reservation_date <- paste(nd_no_show$arrival_date_year,
nd_no_show$arrival_date_month, nd_no_show$arrival_date_day_of_month)
last_update <- as.Date(reservation_date, format = "%Y %B %d") +
nd_no_show$stays_in_weekend_nights + nd_no_show$stays_in_week_nights

temp_hb[no_show_bool, ]$reservation_status_date <- last_update

#Update reservation_status == Canceled
canceled_bool <- is.na(hb_data$reservation_status_date) &
hb_data$reservation_status == 'Canceled'
temp_hb <- temp_hb %>% fill(reservation_status_date)

update_status <-
temp_hb[is.na(temp_hb$reservation_status_date),]$reservation_status
bp <- barplot(table(update_status)) + theme_bw()
hb_data <- temp_hb
```



arrival_date_year

Primero, visualizamos cuales son las filas que tienen este dato faltantes en sus valores. Como son pocos valores podemos hacer un cambio simple donde el año de llegada será el mismo que el año de reserva.

```
hb_nadate_year <- hb_data[is.na(hb_data$arrival_date_year ),]
hb_nadate_year
```

```
hb_data[is.na(hb_data$arrival_date_year),]$arrival_date_year <-
format(hb_nadate_year$reservation_status, format = "%Y")
```

total_of_special_requests <int>	reservation_status <fctr>	reservation_status_date <date>
0	Check-Out	2016-04-03
0	Canceled	2016-06-17
2	Canceled	2017-01-17
0	Canceled	2017-02-28
0	Canceled	2017-02-11
0	Canceled	2017-03-11

Identificación de datos atípicos

Los datos que difieren mucho de los

Code in R:

```
fix_outliers <- function(x, removeNA = TRUE){ #Calculamos los
quantiles 1) por arriba del 5% y por debajo del 95%
  quantiles <- quantile(x, c(0.05, 0.95), na.rm = removeNA)
  x[x<quantiles[1]] <- mean(x, na.rm = removeNA)
  x[x>quantiles[2]] <- median(x, na.rm = removeNA)
  x
}
#                  #
#Variable:Adults   #
#                  #
summary(hotel$adults)
hotel.adults_sin_out <- fix_outliers(hotel$adults)
hist(hotel.adults_sin_out)
boxplot.stats(hotel.adults_sin_out)
```

Plot:

```
> summary(hotel$adults)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   2.000   2.000   1.856   2.000  55.000      12
>
```

Como se puede observar en el summary del atributo Adults del dataset, hay un máximo de 55 mientras la media es 2 , lo que nos hace sospechar de posibles, también podríamos comprobarlo mediante gráficas como el Histograma y el Boxplot, siendo este último el más adecuado, una vez identificamos los valores atípicos metemos a la función “fix_outliers” cada variable que cumpla con esta condición, lo que limpiara el set.
En el código se aprecia de cómo se hace este proceso para una variable, en el repositorio está una versión más extensa con cada variable que sigue el mismo algoritmo solo cambiando la variable que cumpla esta condición

VISUALIZAR DATOS

a. ¿Cuántas reservas se realizan por tipo de hotel? o ¿Qué tipo de hotel prefiere la gente?

```
reserva <- factor(hotel$is_canceled)
ggplot(hotel, aes(x=reserva, fill = hotel)) +
  geom_bar(position = position_dodge()) +
  labs(title = "Reservas por hotel",
       x = "Se cancelaron?",
       y = "Reservas") + theme_bw()
```

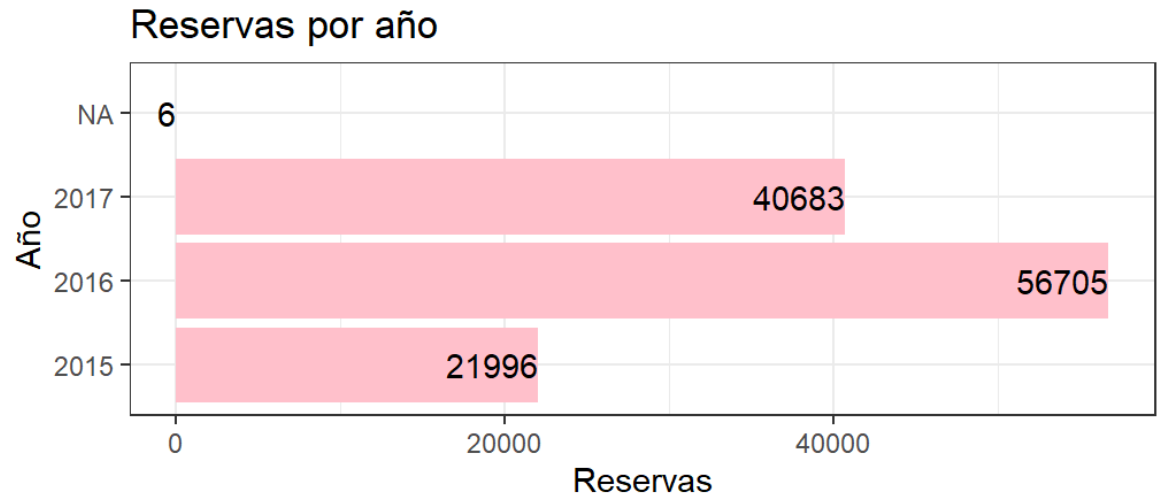
Plot:

Description:La cantidad de cancelaciones de reserva fueron menores en City Hotel, y también fueron en donde más reservas se llegaron a concretar con más de 40000 reservas sin cancelar.

b. ¿Está aumentando la demanda con el tiempo?

```
# Reorganizar mes correctamente
hotel$arrival_date_month <-
  factor(hotel$arrival_date_year)

# Mostrar
ggplot(data = hotel, aes(x = arrival_date_month)) +
  geom_bar(fill = "pink") +
  geom_text(stat = "count", aes(label = ..count..), hjust = 1) +
  coord_flip() + labs(title = "Reservas por año",
                      x = "Año",
                      y = "Reservas") + theme_bw()
```

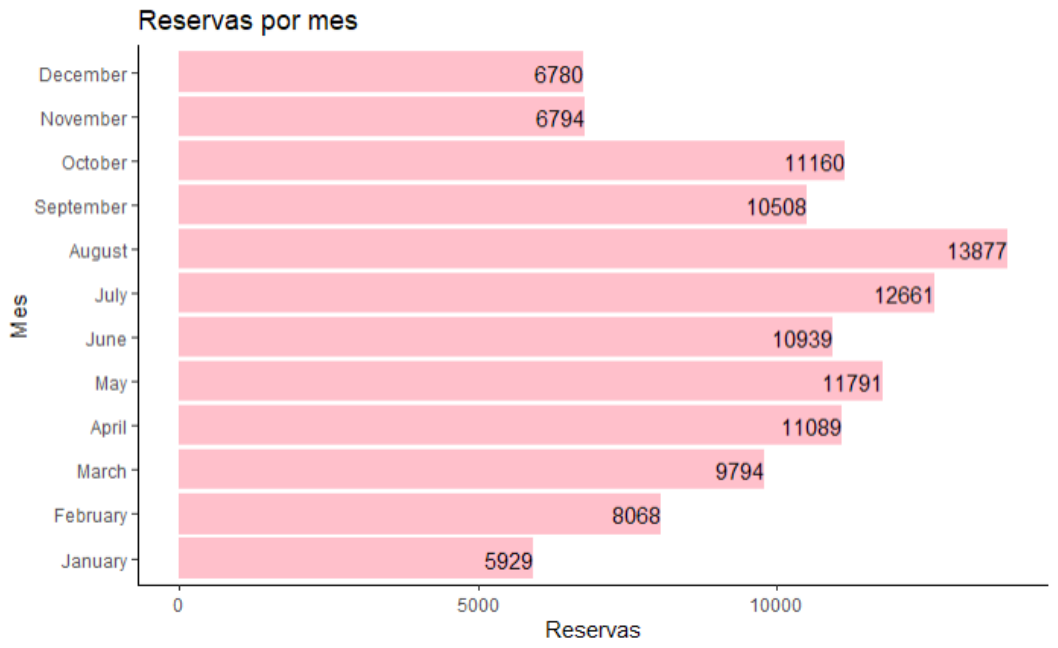
Plot:



Description: Se puede observar que la tendencia no está a la alza si lo ordenamos por año, de hecho tuvo un descenso en el año 2017.

c. ¿Cuándo se producen las temporadas de reservas: alta, media y baja?

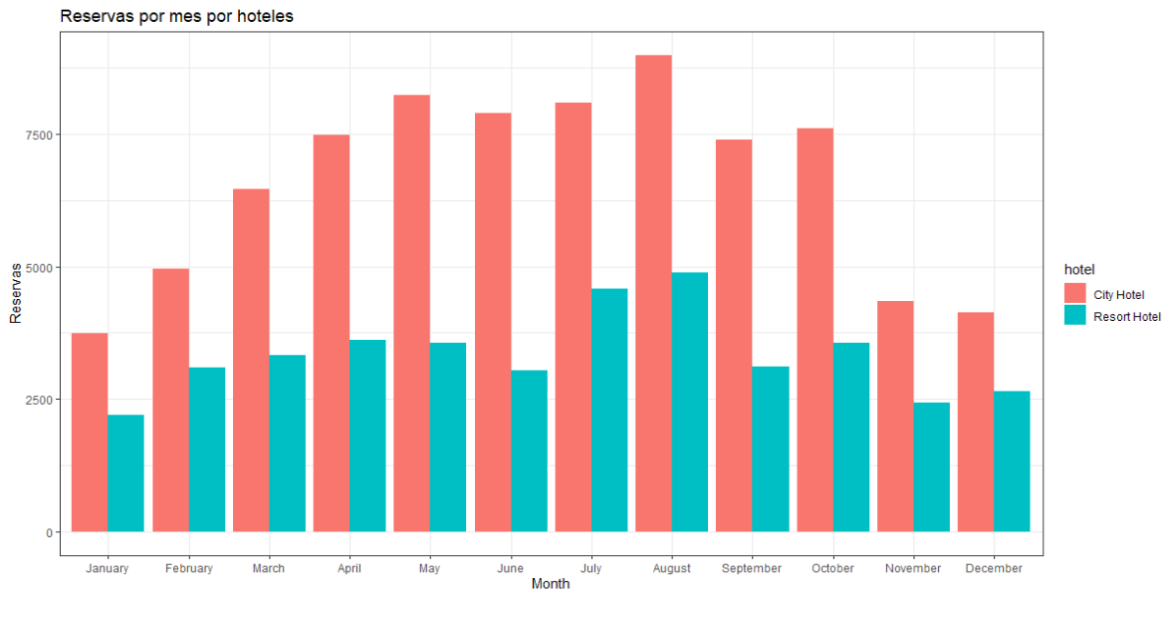
```
# Reorganizar mes correctamente
hb_data$arrival_date_month <-
  factor(hb_data$arrival_date_month, levels = month.name)

# Mostrar
ggplot(data = hb_data, aes(x = arrival_date_month)) +
  geom_bar(fill = "pink") +
  geom_text(stat = "count", aes(label = ..count..), hjust = 1) +
  coord_flip() + labs(title = "Reservas por mes",
                      x = "Mes",
                      y = "Reservas") +
  theme_classic()
```



Las temporadas de reservas más bajas son durante los meses de Noviembre a Febrero, media serían entre los meses de Marzo a Junio y las temporadas altas sería durante los meses de Julio a Octubre. Los datos son hoteles en Portugal siendo estos meses los de verano y de vacaciones, esta pueda ser una de las causas por este aumento en reservas.

d. ¿Cuándo es menor la demanda de reservas y en qué tipo de hotel?

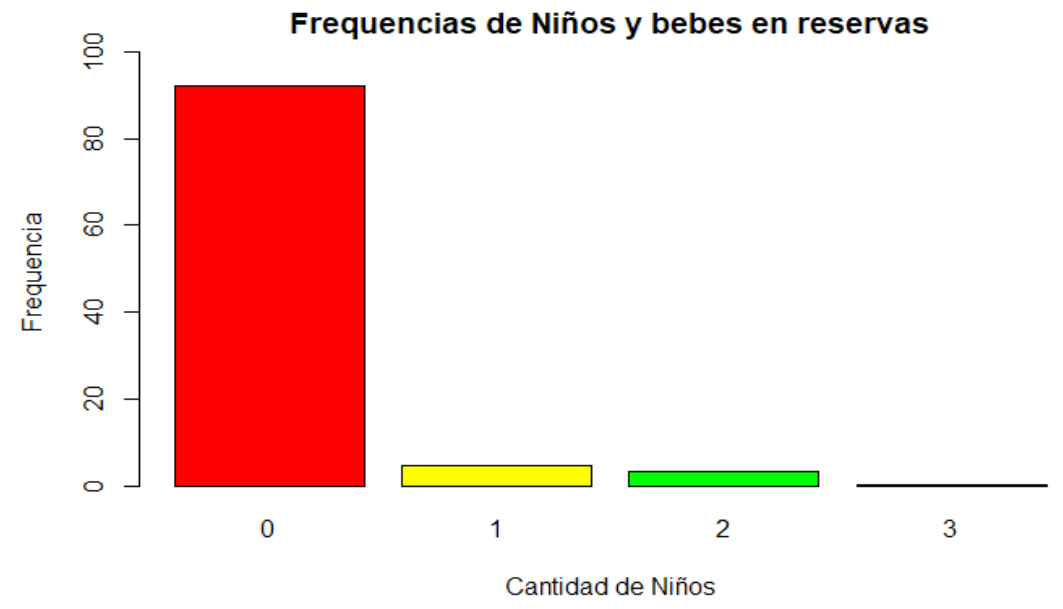
```
ggplot(hb_data, aes(arrival_date_month, fill = hotel)) +
  geom_bar(position = position_dodge()) +
  labs(title = "Reservas por mes por hoteles",
       x = "Month",
       y = "Reservas") + theme_bw()
```

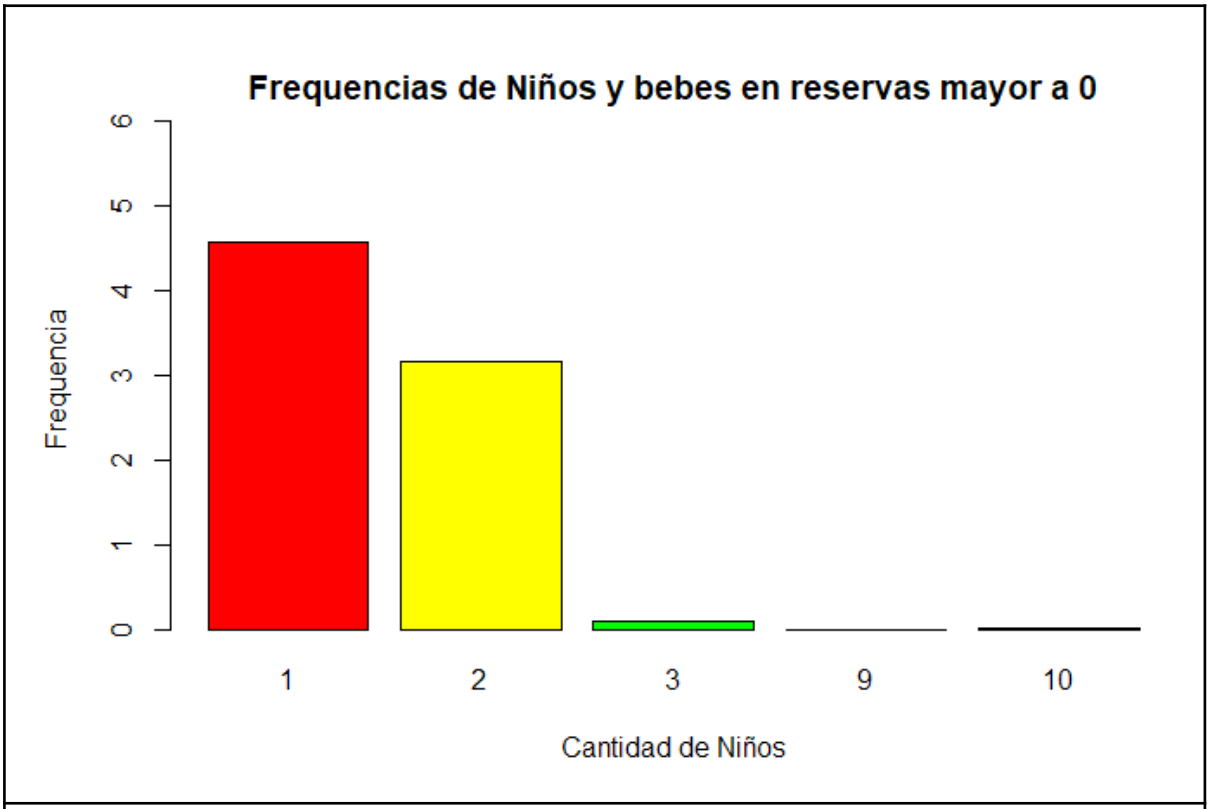


Los meses de menor demanda tanto para el hotel en la ciudad como el hotel en el resort son en el mes de enero y noviembre. Al igual que en la conclusión anterior, el clima en estos meses puede ser un gran factor para contar con un menor cantidad de huéspedes durante estas temporadas.

e. ¿Con qué frecuencia las reservas incluyen niños y/o bebes?

```
reserve_children <- hb_data$children
reserve_babies <- hb_data$babies

reserve_kids <- reserve_children + reserve_babies

kids_table <- table(reserve_kids[reserve_kids < 9])
#Second graph
kids_table_2 <- table(reserve_kids[reserve_kids != 0])

bp <- barplot((kids_table * 100) / nrow(hb_data),
        main="Niños y bebes en reservas",
        xlab= "Kids",
        col=rainbow(6),
        ylim=c(0,100),
        ylab="Frequencia de niños/bebe ",
        beside=TRUE
    ) + theme_bw()

bp <- barplot((kids_table_2 * 100) / nrow(hb_data),
        main="Frecuencias de Niños y bebes en reservas mayor a 0",
        xlab= "Cantidad de Niños",
        col=rainbow(6),
        ylim=c(0,6),
        ylab="Frecuencia ",
        beside=TRUE
    ) + theme_bw()
```

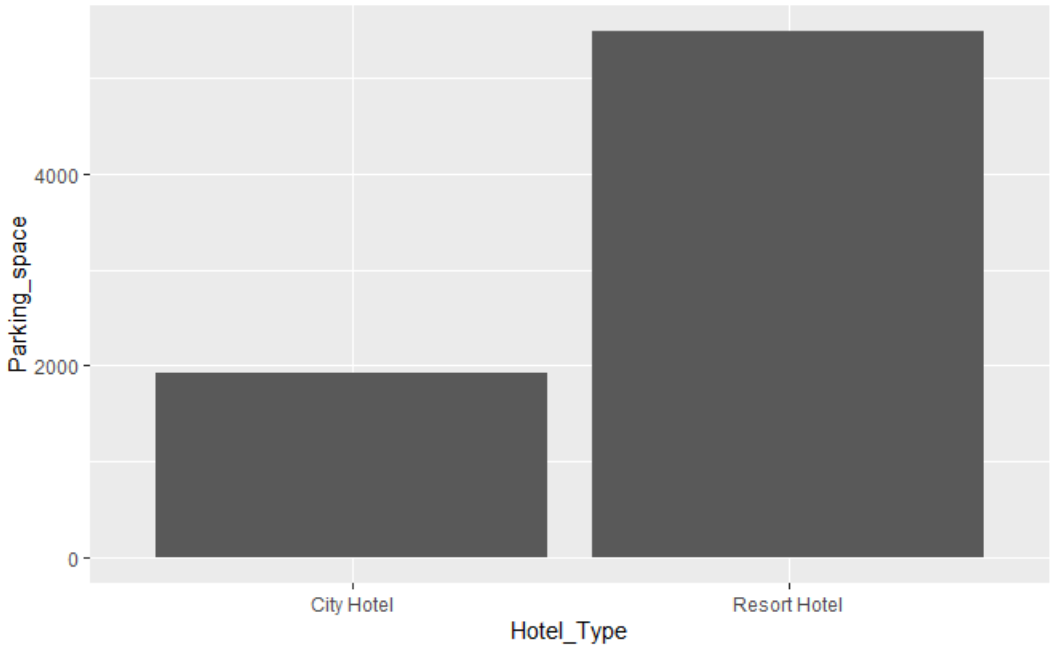




Es evidentemente, que una cantidad significativa de las reservas no cuentan niños o bebes. Es decir, que en estos hoteles las reservas suelen ser para adultos y aun cuando las reservas incluyen niños estas son con dos niños o menos. Es decir, estos hoteles no tienen a las familias con hijos como principal clientes.

f. ¿Es importante contar con espacios de estacionamiento?

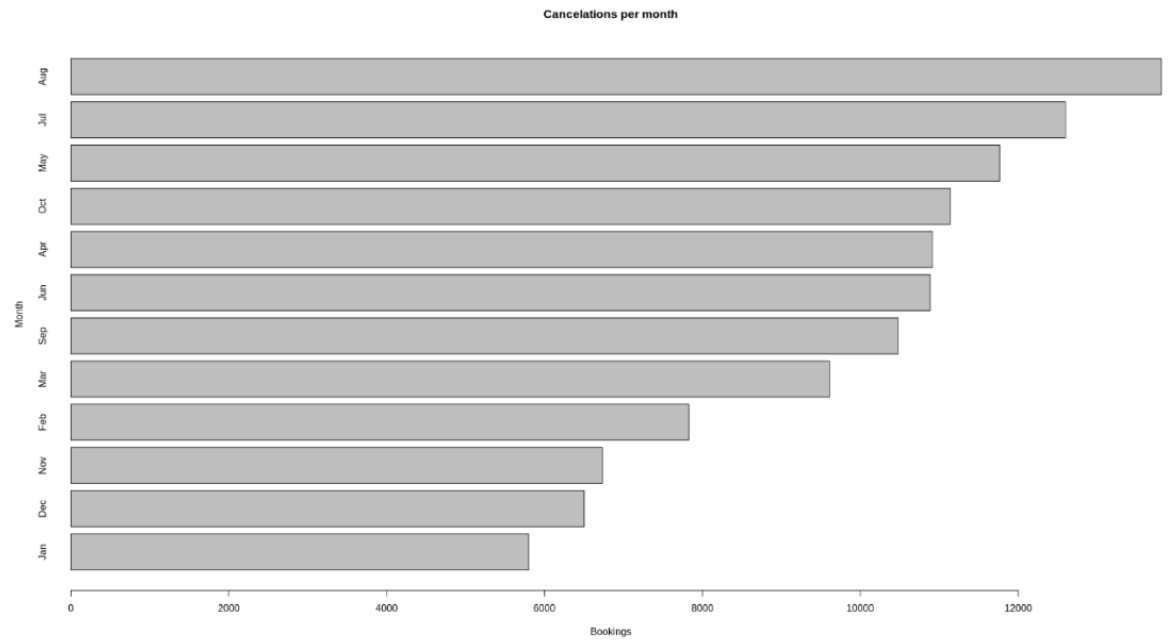
```
parking_space_hotel1 <- hb_data%>% filter(hotel == 'Resort Hotel' &
required_car_parking_spaces > 0 )%>% summarise(n())
parking_space_hotel2 <- hb_data%>% filter(hotel == 'City Hotel' &
required_car_parking_spaces > 0 )%>% summarise(n())
parking <- tribble(
  ~Hotel_Type,     ~Parking_space,
  "Resort Hotel" ,  5490,
  "City Hotel" , 1926
)

ggplot(data=parking) + geom_bar(mapping = aes(x = Hotel_Type, y =Parking_space),
stat = "identity")
```



Es importante contar con espacios de estacionamiento , porque según el reporte extraído , gran parte de los usuarios reservan espacios de estacionamiento al momento de reservar su estadía.

g. ¿En qué meses del año se producen más cancelaciones de reservas?

```
cancelled <- subset(hb_data, reservation_status == "Canceled")
cmonths <- sort(substring(hb_data$arrival_date_month, first = 1 ))
barplot(sort(table(cmonths)), horiz=TRUE, main="Cancelations per month",
        xlab="Bookings", ylab="Month")
```



Como se puede apreciar, los meses que se cancelan más reservaciones son en Agosto y Julio, es decir que se realizan más reservaciones en esas fechas.

CONCLUSIONES PRELIMINARES

Es evidente que este dataset tiene información relevante para el sector turismo, tales como en qué temporadas los huéspedes están más interesados en vacacionar, es decir en qué meses se debe enfocar el marketing al igual de quienes son sus principales clientes. El análisis exploratorio de datos nos muestra.
Además es importante recalcar que para hacer una toma de decisiones apropiada es crítico tener unos datos con una buena consistencia y ajenos al ruido, pues gracias a estas estadísticas es las que se hacen toma de decisiones.Es por esta ultima razon que el preprocesado de datos de cualquier dataset es importante, para que los procesos que le subsigue a este tengan una base sólida.

En conclusión, con los datos extraídos del dataset “Hotel Booking” , se obtiene el comportamiento que tiene un usuario al reservar estadía en un hotel y cuales son sus principales demandas en todo el proceso. Asimismo , los usuarios pueden apreciar todo los beneficios que tienen al realizar una reserva , y cuando es el mejor momento y lugar para realizarlo.

ANEXOS

Repositorio de github: https://github.com/thavs-college-repos/ea-2022-1-cc51