

Universidad Peruana de Ciencias Aplicadas



INFORME DEL TRABAJO PARCIAL

CURSO FUNDAMENTOS DE DATA SCIENCE

Carrera de Ciencias de la Computación

Sección: CC53

Alumnos:	
Código	Nombres y apellidos
U202210161	Diego Antonio Salinas Casaico
u202216148	Salvador Diaz Aguirre
U202215375	Ricardo Rafael Rivas Carrillo

Índice

Índice.....	1
1. Caso de Análisis:.....	2
Origen de los datos.....	2
Casos de uso aplicables.....	2
2. Conjunto de datos (dataset).....	3
3. Análisis exploratorio de datos.....	7
• Visualizar datos:.....	11
4. Conclusiones preliminares.....	16
Bibliografía.....	18

1. Caso de Análisis:

Origen de los datos

El conjunto de datos originales proviene del artículo *Hotel booking demand datasets* de los autores Nuno Antonio, Ana de Almeida y Luis Nunes creado en el año 2019. Este artículo de análisis fue creado como contenido de acceso gratuito para su publicación en la página web “ScienceDirect”, dedicada al almacenamiento de una gran base de datos de búsqueda científica, médica, de ingeniería, etc.

El conjunto de datos está formado por otros dos conjuntos bajo la misma estructura de datos de 31 variables cada uno. El primero, son los datos de la demanda de reserva de un hotel de un centro turístico, describiendo 40 060 observaciones. El segundo, los datos de la demanda de reserva de un hotel de una ciudad, con 79 330 observaciones. Al analizar ambos conjuntos de datos, tenemos un total de 119 390 observaciones.

Casos de uso aplicables

El análisis de los datos de la base de datos proporcionada serán importantes para:

- Los educadores: El análisis de este conjunto de datos puede ser muy beneficioso para enseñar la parte práctica del análisis de datos. Además, de ejemplo como segmentación y clasificación.
- Empresas de publicidad: Al hacer un análisis exhaustivo de las necesidades no dichas de los clientes, los analistas pueden entrenar sus algoritmos de predicción y lanzar estrategias y tácticas para el aumento de publicidad enfocada.

- Los Investigadores: Pueden usar esta base de datos para poder entrenar sus Máquinas de aprendizaje autónomo y mejorar su algoritmos de búsqueda/clasificación.
- Dueños de los hoteles: Identificar tendencias estacionales, preferencias de los clientes y áreas de mejora en la experiencia del huésped, tomar decisiones informadas para mejorar la eficiencia operativa y la satisfacción del cliente.

Los resultados del siguiente análisis deberán dar respuesta a las siguientes preguntas:

- i. ¿Cuántas reservas se realizan por tipo de hotel? o ¿Qué tipo de hotel prefiere la gente?
- ii. ¿Está aumentando la demanda con el tiempo?
- iii. ¿Cuándo se producen las temporadas de reservas: alta, media y baja?
- iv. ¿Cuándo es menor la demanda de reservas?
- v. ¿Cuántas reservas incluyen niños y/o bebés?
- vi. ¿Es importante contar con espacios de estacionamiento?
- vii. ¿En qué meses del año se producen más cancelaciones de reservas?

2. Conjunto de datos (dataset)

Utilizaremos el conjunto de datos mencionado anteriormente. A continuación, se muestra una breve descripción de las 31 variables del conjunto de datos. Se usarán las siglas PMS para referir al Property Management System (sistema de administración de propiedad).

Tabla 1

Descripción del conjunto de datos “Hotel booking demand datasets”

Variable	Descripción
hotel	Tipo de hotel donde se hospedaron los clientes. Puede ser City Hotel o Resort Hotel.
is_canceled	Valor que indica si la reserva fue cancelada. 0 - no fue cancelada. 1 - sí fue cancelada.
lead_time	Número de días entre que la reserva fue ingresada al PMS y la fecha que el cliente llegó al hotel.
arrival_date_year	Año que el cliente llegó al hotel.
arrival_date_month	Mes que el cliente llegó al hotel.
arrival_date_week_number	Número de la semana que el cliente llegó al hotel.
arrival_date_day_of_month	Número del mes que el cliente llegó al hotel.
stays_in_weekend_nights	Número de noches de fin de semana (sábado o domingo) que el cliente se quedó o planeaba quedarse.
stays_in_week_nights	Número de noches de día de semana (lunes a viernes) que el cliente se quedó o planeaba quedarse.
adults	Número de adultos.
children	Número de niños.
babies	Número de bebés.
meal	Tipos de comida reservadas. Pueden ser: BB (bed & breakfast (solo desayuno)), HB (half board

Variable	Descripción
	(desayuno y otra comida)), FB (full board (desayuno, almuerzo y cena)), SC (sin comida)/undefined (no lo definieron los usuarios).
country	Países de origen. Representado en el formato ISO 3155-3:2013.
market_segment	Designación de segmento de mercado. Puede ser Online TA*, Offline TA/TO**, grupos, directo, compañía o complementario. *TA: Travel Agents **TO: Tour Operators
distribution_channel	Canal de distribución de reservas. Puede ser por compañía, directo, GDS* , TA/TO o sin definir. *GDS: Global Distribution System (Sistema de distribución global)
is_repeated_guest	Valor que indica si la reserva fue realizada a nombre de un cliente repetido. 1 sí es repetido. 0 no es repetido.
previous_cancellations	Número de reservas canceladas por el cliente antes de la reserva actual.
previous_bookings_not_canceled	Número de reservas anteriores que el cliente no canceló antes de la reserva actual.
reserved_room_type	Código del tipo de la habitación reservada.

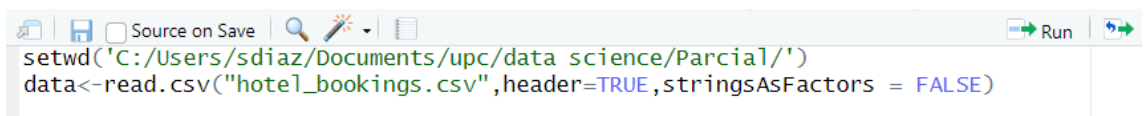
Variable	Descripción
	Alfabéticamente de la A a la J.
assigned_room_type	Código de tipo de la habitación asignada a la reserva. Puede variar del código de reserva por motivos operacionales del hotel. Alfabéticamente de la A a la L.
booking_changes	Número de cambios hechos a la reserva desde el momento de su ingreso al PMS hasta la cancelación o Check-in.
deposit_type	Indicación de si el cliente hizo un depósito para garantizar la reserva. No hubo depósito, no reembolsable (depósito total o mayor al costo de la reserva) o reembolsable.
agent	ID de la agencia de viaje que hizo la reserva.
company	ID de la compañía que hizo la reserva o fue responsable de pagarla.
days_in_waiting_list	Número de días que la reserva estuvo en la lista de espera antes de ser confirmada por el comprador.
customer_type	Tipo de reserva. Puede ser por contrato, en grupo, transitorio (ninguno de los anteriores) o Grupo-Transitorio.
adr	Average Daily Rate (suma de las transacciones de reserva)/(número de noches reservadas)

Variable	Descripción
required_car_parking_spaces	Número de estacionamientos para carro requeridos por cliente.
total_of_special_requests	Números de peticiones especiales hechas por los clientes (cama doble, piso alto, etc.)
reservation_status	Último estado de la reserva. Puede ser: Cancelado, Check-out (hizo el check-in pero ya se fue), No-Show (no hizo el check-in).
reservation_status_date	Fecha que el último estado de reserva fue realizado.

3. Análisis exploratorio de datos

- **Carga de datos**

La carga del dataset deberá considerar los parámetros `header = TRUE`, `stringsAsFactors = FALSE`



```
setwd('C:/Users/sdiaz/Documents/upc/data science/Parcial/')
data<-read.csv("hotel_bookings.csv",header=TRUE,stringsAsFactors = FALSE)
```

- **Inspeccionar datos**

Exploramos los datos del dataset verificando, por ejemplo, estructura, tipo, valores de los datos, nombre de columnas, etc.

Para este análisis, vimos conveniente la conversión de muchos de los datos.


```

$ hotel : chr "Resort Hotel" "Resort Hotel" "Resort Hotel" "Resort Hotel" ...
$ is_canceled : int 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 ...
$ arrival_date_month : chr "July" "July" "July" "July" ...
$ arrival_date_week_number : int 27 27 27 27 27 27 27 27 27 ...
$ arrival_date_day_of_month : int 1 1 1 1 1 1 1 1 1 ...
$ stays_in_weekend_nights : int 0 0 0 0 0 0 0 0 0 ...
$ stays_in_week_nights : int 0 0 1 1 2 2 2 3 3 ...
$ adults : int 2 2 1 1 2 2 2 2 2 ...
$ children : int 0 0 0 0 0 0 0 0 0 ...
$ babies : int 0 0 0 0 0 0 0 0 0 ...
$ meal : chr "BB" "BB" "BB" "BB" ...
$ country : chr "PRT" "PRT" "GBR" "GBR" ...
$ market_segment : chr "Direct" "Direct" "Direct" "Corporate" ...
$ distribution_channel : chr "Direct" "Direct" "Direct" "Corporate" ...
$ is_repeated_guest : int 0 0 0 0 0 0 0 0 0 ...
$ previous_cancellations : int 0 0 0 0 0 0 0 0 0 ...
$ previous_bookings_not_canceled : int 0 0 0 0 0 0 0 0 0 ...
$ reserved_room_type : chr "C" "C" "A" "A" ...
$ assigned_room_type : chr "C" "C" "C" "A" ...
$ booking_changes : int 3 4 0 0 0 0 0 0 0 ...
$ deposit_type : chr "No Deposit" "No Deposit" "No Deposit" "No Deposit" ...
$ agent : chr "NULL" "NULL" "NULL" "304" ...
$ company : chr "NULL" "NULL" "NULL" "NULL" ...
$ days_in_waiting_list : int 0 0 0 0 0 0 0 0 0 ...
$ customer_type : chr "Transient" "Transient" "Transient" "Transient" ...
$ adr : num 0 0 75 75 98 ...
$ required_car_parking_spaces : int 0 0 0 0 0 0 0 0 0 ...
$ total_of_special_requests : int 0 0 0 0 1 1 0 1 0 ...
$ reservation_status : chr "Check-out" "Check-out" "Check-out" "Check-out" ...
$ reservation_status_date : chr "2015-07-01" "2015-07-01" "2015-07-02" "2015-07-02" ...

```

Los datos de tipo chr y otros tipo int fueron convertidos a tipo factor, para una mejor manipulación.

```

data$hotel<-as.factor(data$hotel)
data$is_canceled<-as.factor(data$is_canceled)
data$meal<-as.factor(data$meal)
data$country<-as.factor(data$country)
data$is_repeated_guest<-as.factor(data$is_repeated_guest)
data$arrival_date_month<-as.factor(data$arrival_date_month)
data$market_segment<-as.factor(data$market_segment)
data$distribution_channel<-as.factor(data$distribution_channel)
data$reserved_room_type<-as.factor(data$reserved_room_type)
data$assigned_room_type<-as.factor(data$assigned_room_type)
data$deposit_type<-as.factor(data$deposit_type)
data$agent<-as.factor(data$agent)
data$company<-as.factor(data$company)
data$customer_type<-as.factor(data$customer_type)
data$reservation_status<-as.factor(data$reservation_status)
data$reservation_status_date<-as.Date(data$reservation_status_date)

```

```

      hotel      is_canceled      lead_time      arrival_date_year      arrival_date_month
City Hotel:79330      0:75166      Min. : 0      Min. :2015      August :13877
Resort Hotel:40060      1:44224      1st Qu.: 18      1st Qu.:2016      July   :12661
                                   Median : 69      Median :2016      May    :11791
                                   Mean   :104      Mean   :2016      October:11160
                                   3rd Qu.:160      3rd Qu.:2017      April  :11089
                                   Max.   :737      Max.   :2017      June   :10939
                                   (Other):47873
arrival_date_week_number arrival_date_day_of_month stays_in_weekend_nights stays_in_week_nights
Min. : 1.00      Min. : 1.0      Min. : 0.0000      Min. : 0.0
1st Qu.:16.00      1st Qu.: 8.0      1st Qu.: 0.0000      1st Qu.: 1.0
Median :28.00      Median :16.0      Median : 1.0000      Median : 2.0
Mean :27.17      Mean :15.8      Mean : 0.9276      Mean : 2.5
3rd Qu.:38.00      3rd Qu.:23.0      3rd Qu.: 2.0000      3rd Qu.: 3.0
Max. :53.00      Max. :31.0      Max. :19.0000      Max. :50.0

      adults      children      babies      meal      country
Min. : 0.000      Min. : 0.0000      Min. : 0.000000      BB :92310      PRT :48590
1st Qu.: 2.000      1st Qu.: 0.0000      1st Qu.: 0.000000      FB : 798      GBR :12129
Median : 2.000      Median : 0.0000      Median : 0.000000      HB :14463      FRA :10415
Mean : 1.856      Mean : 0.1039      Mean : 0.007949      SC :10650      ESP : 8568
3rd Qu.: 2.000      3rd Qu.: 0.0000      3rd Qu.: 0.000000      Undefined: 1169      DEU : 7287
Max. :55.000      Max. :10.0000      Max. :10.000000      (Other):28635
                                   NA's :4
      market_segment      distribution_channel      is_repeated_guest      previous_cancellations
Online TA :56477      Corporate: 6677      0:115580      Min. : 0.00000
Offline TA/TO:24219      Direct :14645      1: 3810      1st Qu.: 0.00000
Groups :19811      GDS : 193      Median : 0.00000
Direct :12606      TA/TO :97870      Mean : 0.08712
Corporate : 5295      Undefined: 5      3rd Qu.: 0.00000
Complementary: 743      Max :26.00000

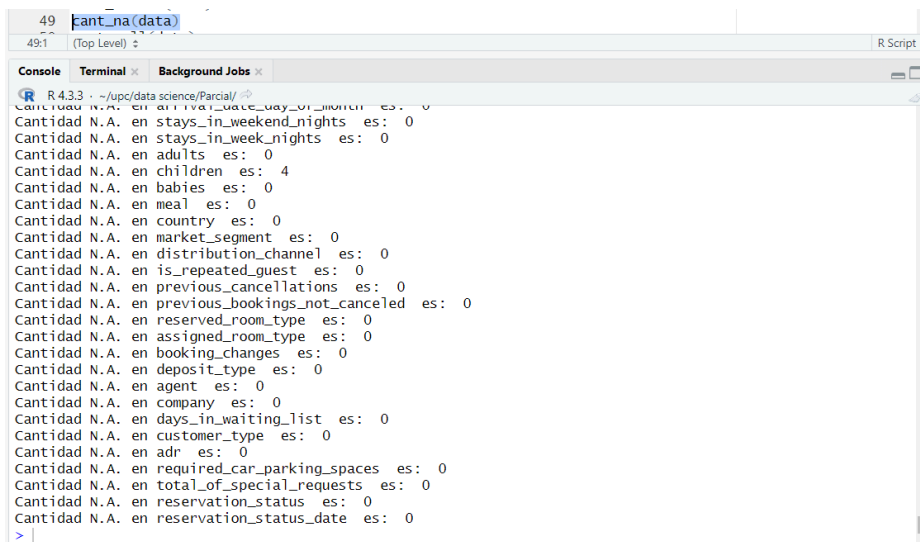
```

- Pre-procesar datos

Valores N.A.: Usamos una función para identificar los valores N.A. del dataset:

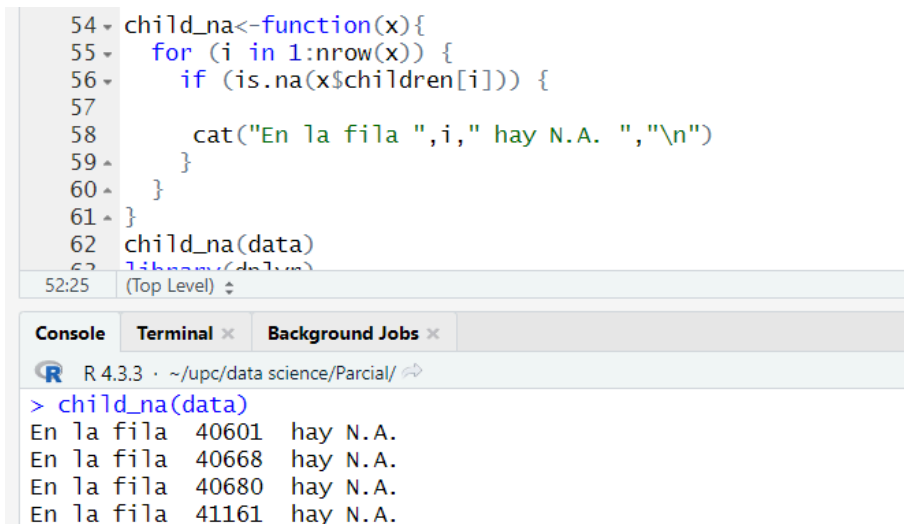
```
cant_na<-function(x){  
  for (i in 1:ncol(x)) {  
    cat("Cantidad N.A. en", colnames(x[i]), " es: ",sum(is.na(x[i])), "\n")  
  }  
}
```

Tras lo cual, vimos que sólo en la columna “children” habían valores N.A.



```
R 4.3.3 ~ /upc/data science/Parcial/  
Cantidad N.A. en arrival_date_day_of_month es: 0  
Cantidad N.A. en stays_in_weekend_nights es: 0  
Cantidad N.A. en stays_in_week_nights es: 0  
Cantidad N.A. en adults es: 0  
Cantidad N.A. en children es: 4  
Cantidad N.A. en babies es: 0  
Cantidad N.A. en meal es: 0  
Cantidad N.A. en country es: 0  
Cantidad N.A. en market_segment es: 0  
Cantidad N.A. en distribution_channel es: 0  
Cantidad N.A. en is_repeated_guest es: 0  
Cantidad N.A. en previous_cancellations es: 0  
Cantidad N.A. en previous_bookings_not_canceled es: 0  
Cantidad N.A. en reserved_room_type es: 0  
Cantidad N.A. en assigned_room_type es: 0  
Cantidad N.A. en booking_changes es: 0  
Cantidad N.A. en deposit_type es: 0  
Cantidad N.A. en agent es: 0  
Cantidad N.A. en company es: 0  
Cantidad N.A. en days_in_waiting_list es: 0  
Cantidad N.A. en customer_type es: 0  
Cantidad N.A. en adr es: 0  
Cantidad N.A. en required_car_parking_spaces es: 0  
Cantidad N.A. en total_of_special_requests es: 0  
Cantidad N.A. en reservation_status es: 0  
Cantidad N.A. en reservation_status_date es: 0  
>
```

Usamos una función para reconocer las filas N.A. de dicha columna:



```
54 child_na<-function(x){  
55   for (i in 1:nrow(x)) {  
56     if (is.na(x$children[i])) {  
57  
58       cat("En la fila ",i," hay N.A. ", "\n")  
59     }  
60   }  
61 }  
62 child_na(data)  
63 library(dplyr)  
52:25 (Top Level)  
Console Terminal Background Jobs  
R 4.3.3 ~ /upc/data science/Parcial/  
> child_na(data)  
En la fila 40601 hay N.A.  
En la fila 40668 hay N.A.  
En la fila 40680 hay N.A.  
En la fila 41161 hay N.A.
```

Al ser una cantidad ínfima de datos comparado al total, decidimos eliminarlo para facilitar las siguientes operaciones. Para esto, usamos la librería “(dplyr)”:

```
new_data<-data%>%filter(!is.na(data$children))
```

Outliers: Identificamos dos casos de valores atípicos. Revisando el resumen de los datos, nos percatamos de diferencias entre mediana y media, un indicador de datos atípicos.

lead_time	adr
Min. : 0	Min. : -6.38
1st Qu.: 18	1st Qu.: 69.29
Median : 69	Median : 94.58
Mean : 104	Mean : 101.83
3rd Qu.: 160	3rd Qu.: 126.00
Max. : 737	Max. : 5400.00

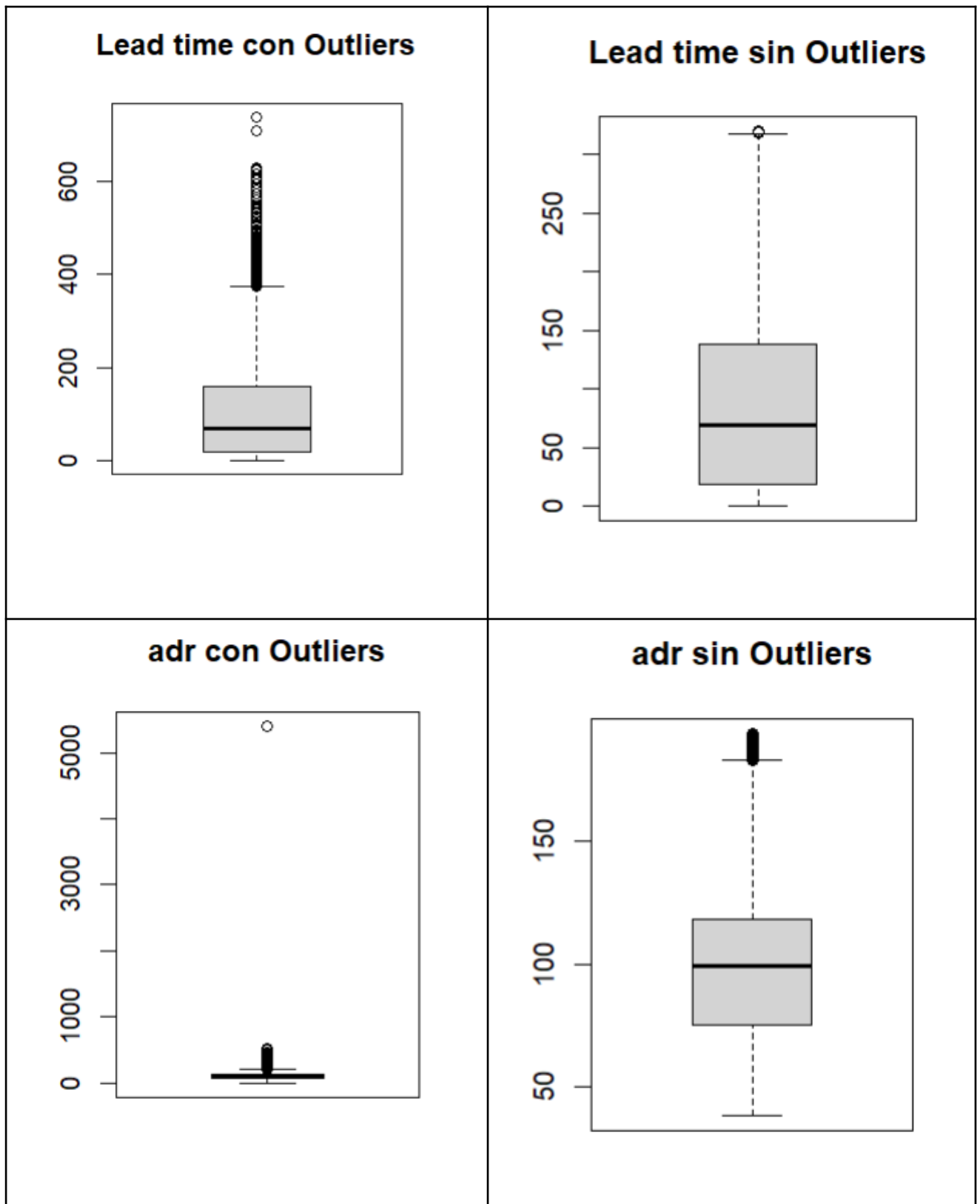
Usando los boxplot, creamos un arreglo que almacena los valores outliers. De esta forma podemos visualizar dichos datos si fuera necesario. Luego, implementamos la función “fix_outliers”, calculando dos percentiles. Si los valores están debajo del quinto percentil, se reemplaza el valor por la media; si están debajo del percentil 95 se reemplaza por la mediana.

```
outlier_values <- boxplot.stats(new_data$lead_time)$out # outlier values.
print(outlier_values)
fix_outliers <- function(x, removeNA = TRUE){
  #Calculamos los cuantiles 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
}
```

Comparación final:

Antes	Después
-------	---------



- Visualizar datos:
 1. Primero, hemos usado solamente la variable “hotel” para saber la cantidad de personas que reservaron cada hotel. Sin embargo, vimos necesario filtrar los

datos que fueron cancelados:

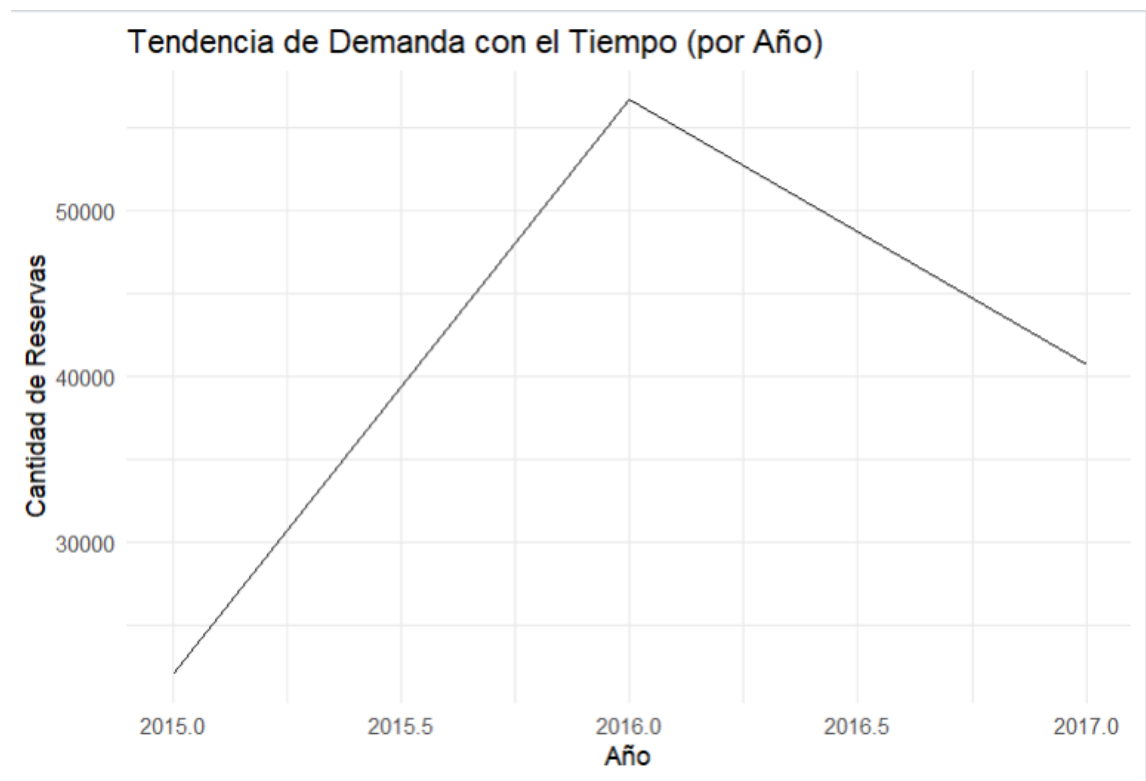
```
> # Filtrar solo las filas no canceladas
> reservas_activas <- new_data1[new_data1$is_canceled == 0,]
> resumen_reservas <- summarise(group_by(reservas_activas, hotel),
+                               Cantidad_Reservas = n())
> print(resumen_reservas)
# A tibble: 2 × 2
  hotel      Cantidad_Reservas
  <fct>          <int>
1 City Hotel      46228
2 Resort Hotel    28938
```

Luego, usamos un gráfico de barras para una mejor visualización:

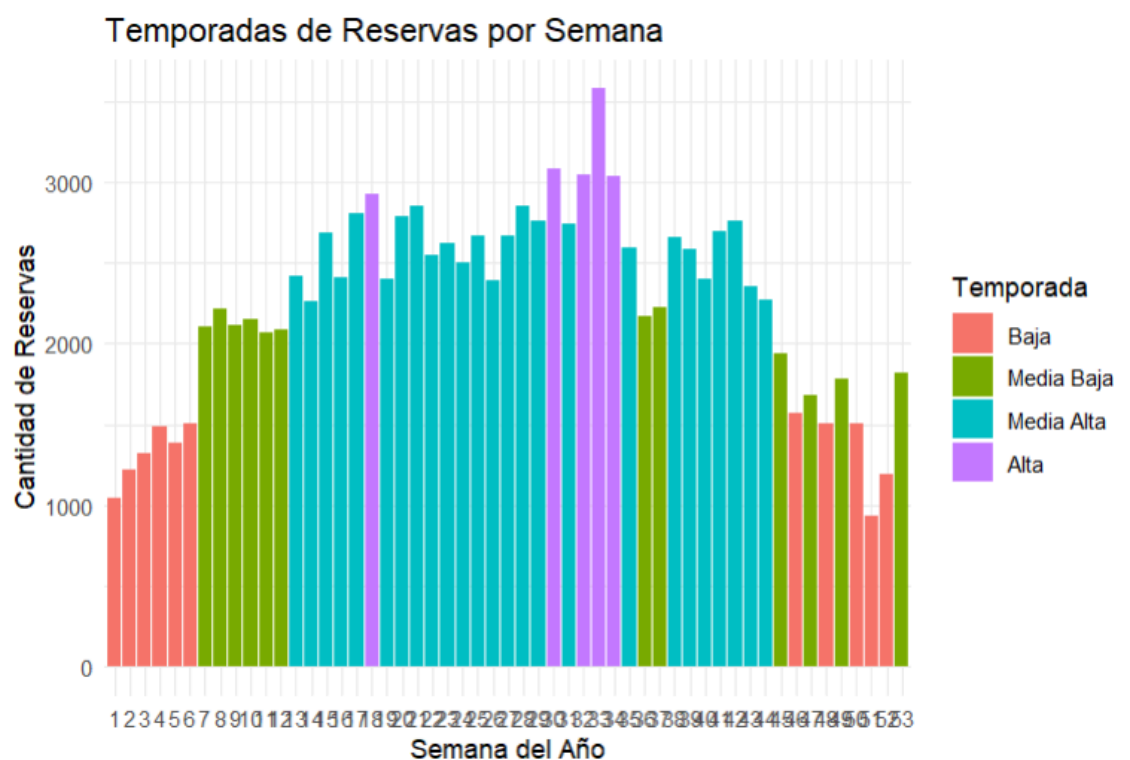


2. Para saber la tendencia de la demanda con el paso del tiempo, usamos los años. Primero, convertimos los datos a numérico para una mejor manipulación. Con esto, pudimos hacer un gráfico de línea que grafica la tendencia con el paso de

los años.



- Para saber las temporadas de reservas (alta, media, baja) hemos utilizado la variable del número de la semana, ya que es la más cercana para saber las ocurrencias de todo un año.



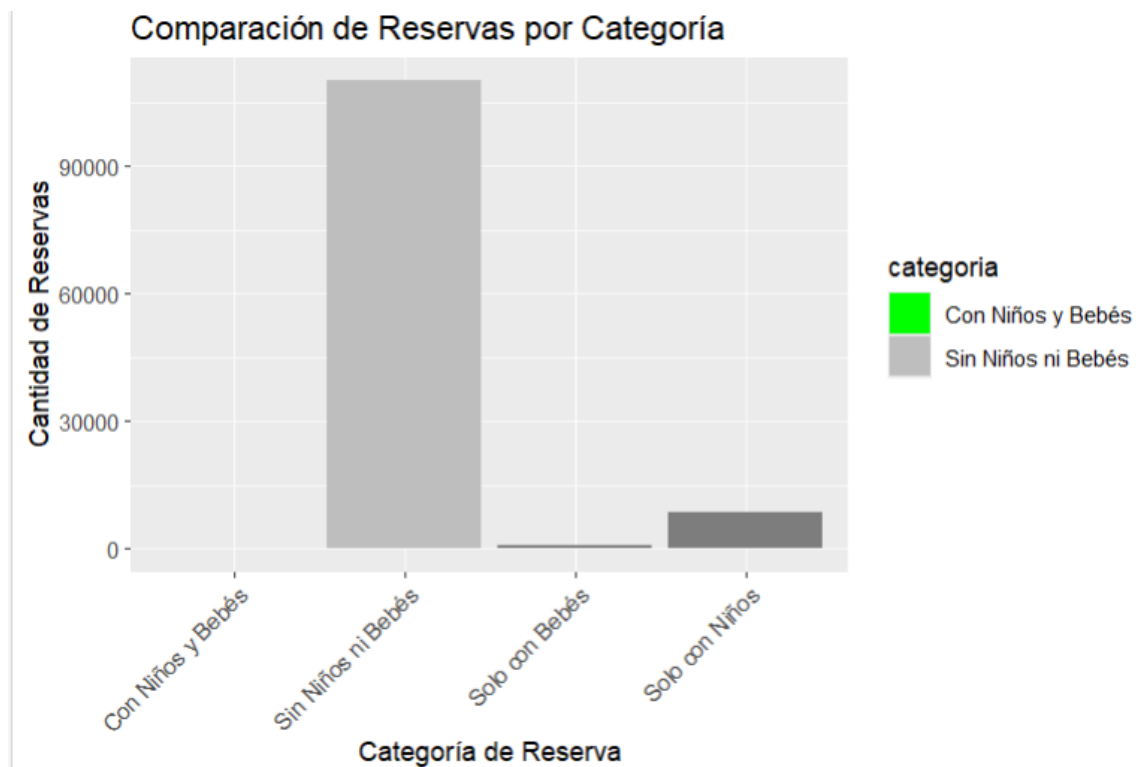
4. A

5. Para saber la cantidad de niños y bebés que habían en cada reserva, primero

usamos una función sum para tener el número exacto:

```
reservas_con_ninos <- sum(clean_data$children > 0) +  
  sum(clean_data$babies > 0)  
  
cat("El número de reservas que incluyen niños y/o bebés es:", reservas_con_ninos, "\n")
```

Usamos un gráfico para visualizar la comparación:



6. Visualizamos la cantidad de personas que requieren parking, debido a que nos

interesaba saber si era realmente importante para los hoteles contar con un

espacio de estacionamiento. Primero, quisimos saber el porcentaje para tener un

valor exacto del cual darnos una idea.

```
# Pregunta vi: ¿Es importante contar con espacios de estacionamiento?  
espacios_estacionamiento <- sum(clean_data$required_car_parking_spaces > 0)  
porcentaje_con_estacionamiento <- (espacios_estacionamiento / nrow(clean_data)) * 100  
cat("El", porcentaje_con_estacionamiento, "% de las reservas requieren espacios de estacionamiento.\n")
```

Del anterior código, tenemos el resultado que el 6.21% de personas no

requieren espacios de estacionamiento.

A continuación, graficamos para demostrar más fácilmente la cantidad de personas:



7. En la siguiente visualización, hemos utilizado “arrival_date_month” y “is_canceled” para saber en qué meses han habido más cancelaciones. Hemos creado una tabla donde hemos ordenado de mayor a menor y mostrado los tres primeros meses, para así poder visualizar los meses con más cancelaciones:

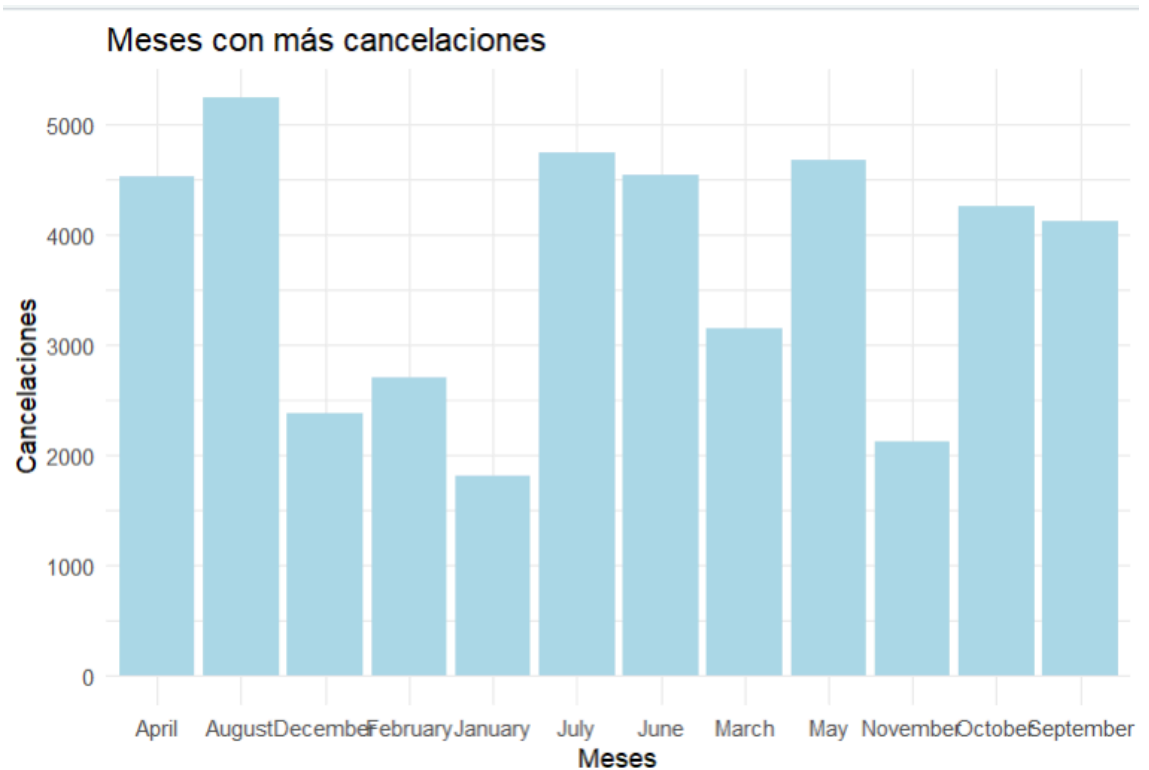
```
# Filtrar las cancelaciones
cancelaciones_por_mes <- clean_data %>%
  filter(is_canceled == 1) %>%
  group_by(arrival_date_month) %>%
  summarise(cancelaciones = n())

# Ordenando los meses por número de cancelaciones (de mayor a menor)
cancelaciones_por_mes <- cancelaciones_por_mes %>%
  arrange(desc(cancelaciones))

# Mostramos los tres meses con más cancelaciones
cat("Los tres meses con más cancelaciones de reservas son:\n")
print(head(cancelaciones_por_mes, 3))
```

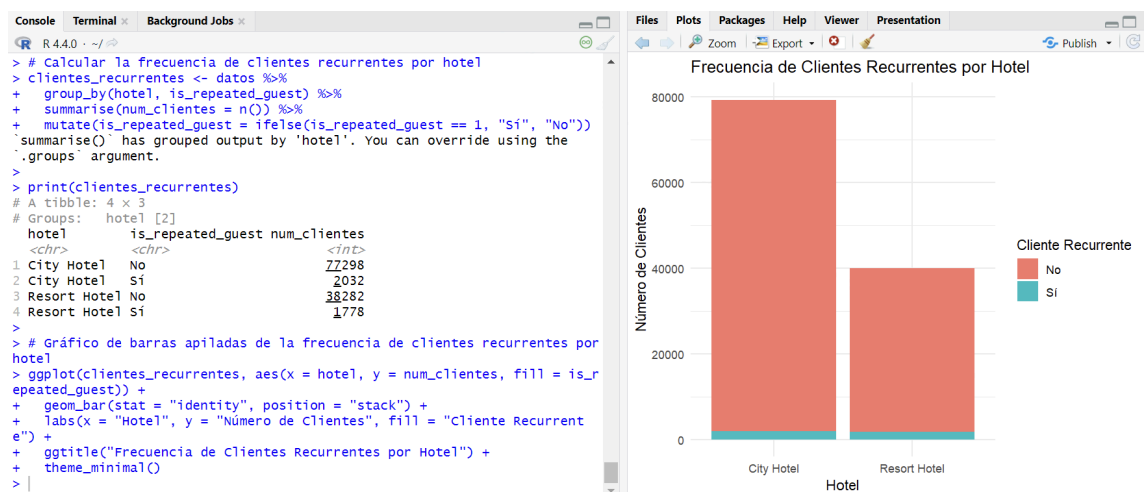
	arrival_date_month	cancelaciones
	<fct>	<int>
1	August	5235
2	July	4742
3	May	4677

Además, hemos creado un gráfico de barras que compara todas las cancelaciones de todos los meses:



Preguntas Extra:

1. Como primer extra hemos pensado en cuántos clientes han regresado a algún hotel, para ello hemos usado “is_repeated_guest” así como “hotel” para poder separar según los hoteles cuantos clientes han regresado.



2. Como segundo extra hemos pensado en el porcentaje de clientes a los cuales les han cambiado de tipo de habitación, con respecto al tipo de habitación que reservaron primeramente, para ello usamos “assigned_room_type” y lo hemos comparado con “reserved_room_type” luego los separamos según el “hotel”. Y así hallamos el porcentaje de cambio de habitaciones según el hotel.



4. Conclusiones preliminares

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

El City Hotel lo prefieren 46228 personas, el Resort Hotel lo prefieren 28938. Teniendo en cuenta sólo las reservas que no fueron canceladas, la gente prefiere el City Hotel.

- b. ¿Está aumentando la demanda con el tiempo?

La demanda aumentó linealmente el primer año. Luego, disminuyó con menor velocidad que cuando subió.

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

Las temporadas más altas ocurren a mitad de año, las más bajas ocurren a

principio y final de año y la media está a la mitad, intercalando también a final de año.

- d. ¿Cuándo es menor la demanda de reservas?

La menor demanda ocurre en la semana 51. Es decir, a mitad de diciembre.

- e. ¿Cuántas reservas incluyen niños y/o bebés?

El número de reservas que incluyen niños y/o bebés es 9507.

- f. ¿Es importante contar con espacios de estacionamiento?

Teniendo en cuenta que el porcentaje de usuarios que necesitaron una dichos espacios fue de 6.21%, a simple vista nos puede parecer un número bastante bajo. Sin embargo, necesitaríamos saber cuántos recursos se gastan en construir los estacionamiento, cuántos se construyen al mismo tiempo, y cuántos estacionamientos no son usados. Aún así, una ganancia del 6% de usuarios puede resultar significativa para un negocio, por lo cual concluimos que sí es importante.

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

Los meses del año con más cancelaciones de reservas son Agosto, Julio y Mayo, con 5235, 4742 y 4677 cancelaciones respectivamente.

Extras:

- a. ¿Cuántos clientes han regresado de nuevo a los hoteles?

En City Hotel han regresado 2032 clientes a reservar habitación más de una vez, y para Resort Hotel han regresado 1778 clientes.

- b. ¿Cuál es el porcentaje de reservas las cuales han experimentado un cambio de habitación respecto a la reserva inicial?

Para City Hotel, un 9,07% de reservas han sufrido de un cambio de habitación, mientras que Resort Hotel ha sufrido de un 19,3% de cambios.

Bibliografía

António, N., Almeida, A., & Nunes, L. (2019). Hotel booking demand datasets. *Data In Brief*, 22, 41-49. <https://doi.org/10.1016/j.dib.2018.11.126>