UPC
Universidad Peruana
de Ciencias Aplicadas

INFORME DE TRABAJO PARCIAL

CC216 – Fundamentos de Data Science

Carrera de Ciencias de la Computación

SECCIÓN: CC53

DOCENTE: Javier Ulises Rosales Huamanchumo

ALUMNOS:

U20201F678 - Joaquin Sebastian Ruiz Ramirez

U202218811 - Andrés Eduardo Carbajal Rojas

U20201F576 - Jimena Alexsandra Quintana Noa

U20221C488 - Freddy Alejandro Cuadros Contreras

U202211514 - Juan Pablo Julian Guija Solis

2024 - 01

Tabla de Contenido

1. Caso de Análisis

1.1. Origen de Datos

El origen de este dataset viene de un Property management system de hoteles de una base de datos en SQL, proveniente del instituto universitario de lisboa, publicado para análisis el 5 de octubre de 2018 y liberado al público desde el 12 de diciembre del mismo año

1.2. Casos Aplicables

Este análisis explora la demanda de dos tipos de hoteles, urbano y resort, utilizando datos pre pandemia para identificar patrones en los perfiles de las personas que hacen reservaciones, la demanda por tipo de habitación y los canales de reserva más utilizados, entre otros aspectos. Al comparar estos datos con tendencias actuales post pandemia, podemos obtener insights valiosos sobre cómo han cambiado las preferencias y comportamientos de los clientes.

Beneficiarios:

Gerentes de Hoteles: Utilizarán este análisis para adaptar sus estrategias comerciales, optimizando aspectos como precios, promociones y servicios para alinearse mejor con las tendencias actuales.

Inversores en el Sector Hotelero: Pueden evaluar la viabilidad de inversiones futuras basadas en tendencias emergentes y la recuperación del mercado.

Agencias de Marketing: Podrían diseñar campañas más efectivas basadas en un entendimiento profundo de los cambios en el mercado turístico.

1.3. Preguntas y visualizaciones

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 bebes?

vi. ¿Es importante contar con espacios de estacionamiento?

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

2. Conjunto de Datos

Por el análisis exploratorio de datos visto en el punto 3, entendemos que cada variable representa lo siguiente:

- **hotel:** Indica el tipo de hotel al que se refiere la reserva. Puede ser "Resort Hotel" o "City Hotel".
- **is_canceled:** Es una variable binaria que indica si la reserva fue cancelada (1) o no (0).
- **lead_time:** Representa el tiempo en días desde la fecha de la reserva hasta la fecha de llegada al hotel. Es una medida de cuánto tiempo antes se realizó la reserva.
- **arrival_date_year:** Representa el año de llegada al hotel.
- **arrival_date_month:** Representa el mes de llegada al hotel.
- **arrival_date_week_number:** Representa el número de semanas del año de llegada al hotel.
- **arrival_date_day_of_month:** Representa el día del mes de llegada al hotel.

- **stays_in_weekend_nights:** Indica el número de noches de estancia durante los fines de semana (sábado o domingo).
- **stays_in_week_nights:** Indica el número de noches de estancia durante los días de semana (de lunes a viernes).
- **adults:** Representa el número de adultos incluidos en la reserva.
- **children:** Representa el número de niños incluidos en la reserva.
- **babies:** Representa el número de bebés incluidos en la reserva.
- **meal:** Indica el tipo de comida reservada, como "BB" (Bed & Breakfast), "HB" (Half Board), "FB" (Full Board), "SC" (Self Catering) o "Undefined" (Sin definir).
- **country:** Representa el país de origen del cliente.
- **market_segment:** Indica el segmento de mercado al que pertenece la reserva, como "TA" (Travel Agents), "TO" (Tour Operators), etc.
- **distribution_channel:** Indica el canal de distribución de la reserva, como "Directo", "Agencia", etc.
- **is_repeated_guest:** Es una variable binaria que indica si el huésped ha realizado reservas previas en el hotel (1) o no (0).
- **previous_cancellations:** Indica el número de reservas previas que fueron canceladas por el cliente antes de la reserva actual.
- **previous_bookings_not_canceled:** Indica el número de reservas previas que no fueron canceladas por el cliente antes de la reserva actual.
- **reserved_room_type:** Representa el tipo de habitación reservada.
- **assigned_room_type:** Representa el tipo de habitación asignada al cliente en el momento del check-in.
- **booking_changes:** Representa el número de cambios realizados a la reserva desde el momento de la reserva hasta el momento del check-in o cancelación.
- **deposit_type:** Indica el tipo de depósito realizado para garantizar la reserva.
- **agent:** Representa el ID de la agencia de viajes que realizó la reserva.
- **company:** Representa el ID de la empresa que realizó la reserva.
- **days_in_waiting_list:** Indica el número de días que la reserva estuvo en lista de espera antes de ser confirmada al cliente.
- **customer_type:** Indica el tipo de cliente, como "Transient", "Contract", "Group", etc.
- **adr (Average Daily Rate):** Representa la tarifa diaria promedio calculada dividiendo la suma de todas las transacciones de alojamiento por el número total de noches de estancia.
- **required_car_parking_spaces:** Indica el número de plazas de aparcamiento requeridas por el cliente.

- **total_of_special_requests:** Indica el número total de peticiones especiales hechas por el cliente.
- **reservation_status:** Indica el último estado de la reserva, como "Canceled", "Check-Out", "No-Show", etc.
- **reservation_status_date:** Representa la fecha en la que se estableció el último estado de la reserva

3. Análisis Exploratorio de Datos

3.1. Carga de Datos

Para iniciar cargamos los datos, con los encabezados de cada columna y dejando los tipo string por defecto.

Aplicamos este script:

```
df_hotel_bookings<-read.csv('hotel_bookings_original.csv'
,header=TRUE,stringsAsFactors = FALSE)
```

3.2. Inspección de Datos

Para ver los tipos de datos de cada variable y el nombre de las columnas, aplicaremos el script str:

```
str(df_hotel_bookings)
```

```
'data.frame':   119390 obs. of  32 variables:
 $ hotel                         : chr  "Resort Hotel" "Resort Hotel" "Resort Hotel" "Resort Hotel" ..
 $ 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            : chr  "July" "July" "July" "July" ...
 $ 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  0 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 0 ...
 $ babies                        : int  0 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 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            : 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 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 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 0 ...
 $ total_of_special_requests     : int  0 0 0 0 1 1 0 1 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" ...
```

Aplicamos scripts adicionales.

```
head(df_hotel_bookings)
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	...	deposit_type	agent	company
	<chr>	<int>	<int>	<int>	<chr>	<int>	<int>	<int>	<int>	<int>	...	<chr>	<chr>	<chr>
1	Resort Hotel	0	342	2015	July	27	1	0	0	2	...	No Deposit	NULL	NULL
2	Resort Hotel	0	737	2015	July	27	1	0	0	2	...	No Deposit	NULL	NULL
3	Resort Hotel	0	7	2015	July	27	1	0	1	1	...	No Deposit	NULL	NULL
4	Resort Hotel	0	13	2015	July	27	1	0	1	1	...	No Deposit	304	NULL
5	Resort Hotel	0	14	2015	July	27	1	0	2	2	...	No Deposit	240	NULL
6	Resort Hotel	0	14	2015	July	27	1	0	2	2	...	No Deposit	240	NULL

```
tail(df_hotel_bookings)
```

A data.frame: 6 × 32

eled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	...	deposit_type	agent	company	days_in_waiting_list	customer_type	adr
int>	<int>	<int>	<chr>	<int>	<int>	<int>	<int>	<int>	...	<chr>	<chr>	<chr>	<int>	<chr>	<dbl>
0	21	2017	August	35	30	2	5	2	...	No Deposit	394	NULL	0	Transient	96.14
0	23	2017	August	35	30	2	5	2	...	No Deposit	394	NULL	0	Transient	96.14
0	102	2017	August	35	31	2	5	3	...	No Deposit	9	NULL	0	Transient	225.43
0	34	2017	August	35	31	2	5	2	...	No Deposit	9	NULL	0	Transient	157.71
0	109	2017	August	35	31	2	5	2	...	No Deposit	89	NULL	0	Transient	104.40
0	205	2017	August	35	29	2	7	2	...	No Deposit	9	NULL	0	Transient	151.20

```
summary(df_hotel_bookings)
```

```
    hotel          is_canceled       lead_time      arrival_date_year
 Length:119390      Min.   :0.0000   Min.   :  0   Min.   :2015
 Class :character   1st Qu.:0.0000   1st Qu.: 18   1st Qu.:2016
 Mode  :character   Median :0.0000   Median : 69   Median :2016
                    Mean   :0.3704   Mean   :104   Mean   :2016
                    3rd Qu.:1.0000   3rd Qu.:160   3rd Qu.:2017
                    Max.   :1.0000   Max.   :737   Max.   :2017

 arrival_date_month arrival_date_week_number arrival_date_day_of_month
 Length:119390      Min.   : 1.00            Min.   : 1.0
 Class :character   1st Qu.:16.00            1st Qu.: 8.0
 Mode  :character   Median :28.00            Median :16.0
                    Mean   :27.17            Mean   :15.8
                    3rd Qu.:38.00            3rd Qu.:23.0
                    Max.   :53.00            Max.   :31.0

 stays_in_weekend_nights stays_in_week_nights     adults
 Min.   : 0.0000         Min.   : 0.0         Min.   : 0.000
 1st Qu.: 0.0000         1st Qu.: 1.0         1st Qu.: 2.000
 Median : 1.0000         Median : 2.0         Median : 2.000
 Mean   : 0.9276         Mean   : 2.5         Mean   : 1.856
 3rd Qu.: 2.0000         3rd Qu.: 3.0         3rd Qu.: 2.000
 Max.   :19.0000         Max.   :50.0         Max.   :55.000

    children           babies              meal             country
 Min.   : 0.0000   Min.   : 0.000000   Length:119390      Length:119390
 1st Qu.: 0.0000   1st Qu.: 0.000000   Class :character   Class :character
 Median : 0.0000   Median : 0.000000   Mode  :character   Mode  :character
 Mean   : 0.1039   Mean   : 0.007949
 3rd Qu.: 0.0000   3rd Qu.: 0.000000
 Max.   :10.0000   Max.   :10.000000
 NA's   :4
 market_segment     distribution_channel is_repeated_guest
 Length:119390      Length:119390        Min.   :0.00000
 Class :character   Class :character     1st Qu.:0.00000
 Mode  :character   Mode  :character     Median :0.00000
                                        Mean   :0.03191
                                        3rd Qu.:0.00000
                                        Max.   :1.00000
```

```
 previous_cancellations previous_bookings_not_canceled reserved_room_type
 Min.   : 0.00000       Min.   : 0.0000                Length:119390
 1st Qu.: 0.00000       1st Qu.: 0.0000                Class :character
 Median : 0.00000       Median : 0.0000                Mode  :character
 Mean   : 0.08712       Mean   : 0.1371
 3rd Qu.: 0.00000       3rd Qu.: 0.0000
 Max.   :26.00000       Max.   :72.0000

 assigned_room_type booking_changes   deposit_type          agent
 Length:119390      Min.   : 0.0000   Length:119390      Length:119390
 Class :character   1st Qu.: 0.0000   Class :character   Class :character
 Mode  :character   Median : 0.0000   Mode  :character   Mode  :character
                    Mean   : 0.2211
                    3rd Qu.: 0.0000
                    Max.   :21.0000

   company          days_in_waiting_list customer_type          adr
 Length:119390      Min.   :  0.000      Length:119390      Min.   :  -6.38
 Class :character   1st Qu.:  0.000      Class :character   1st Qu.:  69.29
 Mode  :character   Median :  0.000      Mode  :character   Median :  94.58
                    Mean   :  2.321                         Mean   : 101.83
                    3rd Qu.:  0.000                         3rd Qu.: 126.00
                    Max.   :391.000                         Max.   :5400.00

 required_car_parking_spaces total_of_special_requests reservation_status
 Min.   :0.00000             Min.   :0.0000            Length:119390
 1st Qu.:0.00000             1st Qu.:0.0000            Class :character
 Median :0.00000             Median :0.0000            Mode  :character
 Mean   :0.06252             Mean   :0.5714
 3rd Qu.:0.00000             3rd Qu.:1.0000
 Max.   :8.00000             Max.   :5.0000

 reservation_status_date
 Length:119390
 Class :character
 Mode  :character
```

Luego de aplicar esos script entendemos que hay variables cualitativas/categóricas como cuantitativas/numéricas.

3.3. Pre-Procesar Datos

En la inspección de datos ejecutamos comandos que nos permitieron ver que algunas variables/columnas contienen datos vacíos o atípicos, lo que dificultará el proceso de análisis para cualquier objetivo, por lo tanto se debe aplicar una estrategias para eliminar o transformar esos datos.
Primero identificamos datos faltantes o NA para lo cual usamos una funcion:

```
sin_valor <- function(x){sum = 0
for(i in 1:ncol(x)){cat("En la columna",colnames(x[i]),"total de valores
NA:",colSums(is.na(x[i])),"\n")}}
sin_valor(df_hotel_bookings)
```

```
En la columna hotel total de valores NA: 0
En la columna is_canceled total de valores NA: 0
En la columna lead_time total de valores NA: 0
En la columna arrival_date_year total de valores NA: 0
En la columna arrival_date_month total de valores NA: 0
En la columna arrival_date_week_number total de valores NA: 0
En la columna arrival_date_day_of_month total de valores NA: 0
En la columna stays_in_weekend_nights total de valores NA: 0
En la columna stays_in_week_nights total de valores NA: 0
En la columna adults total de valores NA: 0
En la columna children total de valores NA: 4
En la columna babies total de valores NA: 0
En la columna meal total de valores NA: 0
En la columna country total de valores NA: 0
En la columna market_segment total de valores NA: 0
En la columna distribution_channel total de valores NA: 0
En la columna is_repeated_guest total de valores NA: 0
En la columna previous_cancellations total de valores NA: 0
En la columna previous_bookings_not_canceled total de valores NA: 0
En la columna reserved_room_type total de valores NA: 0
En la columna assigned_room_type total de valores NA: 0
En la columna booking_changes total de valores NA: 0
En la columna deposit_type total de valores NA: 0
En la columna agent total de valores NA: 0
En la columna company total de valores NA: 0
En la columna days_in_waiting_list total de valores NA: 0
En la columna customer_type total de valores NA: 0
En la columna adr total de valores NA: 0
En la columna required_car_parking_spaces total de valores NA: 0
En la columna total_of_special_requests total de valores NA: 0
En la columna reservation_status total de valores NA: 0
En la columna reservation_status_date total de valores NA: 0
```

Vemos que se tiene 4 valores NA en la columna children, al ser tan pocos podemos eliminar las filas de estos datos.

```
df_hotel_bookings<-na.omit(df_hotel_bookings)
#Usamos summary para verificar que no hay ningun NA en children
```

```
1 df_hotel_bookings<-na.omit(df_hotel_bookings)
2 summary(df_hotel_bookings$children)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.0000  0.0000  0.0000  0.1039  0.0000 10.0000
```

Adicionalmente se encontraron datos atípicos en la columna adr pues esta tiene como valor mínimo un negativo lo cual no tiene sentido ya que esta columna habla del costo de la reserva. Pero no se tomará en cuenta pues no se realizará análisis en torno a esta variable.
Pasamos guardar la data limpia en otro archivo .csv

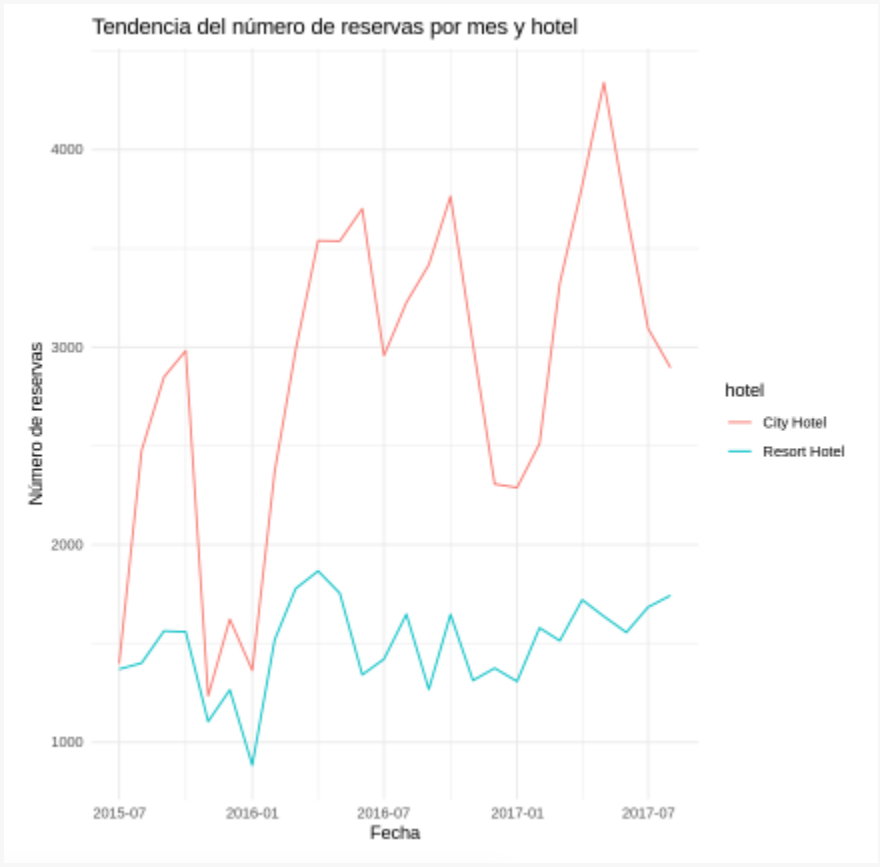
```
1 write.csv(df_hotel_bookings, "hotel_bookings_final.csv", row.names = FALSE)
```

3.4. Visualizar Datos

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

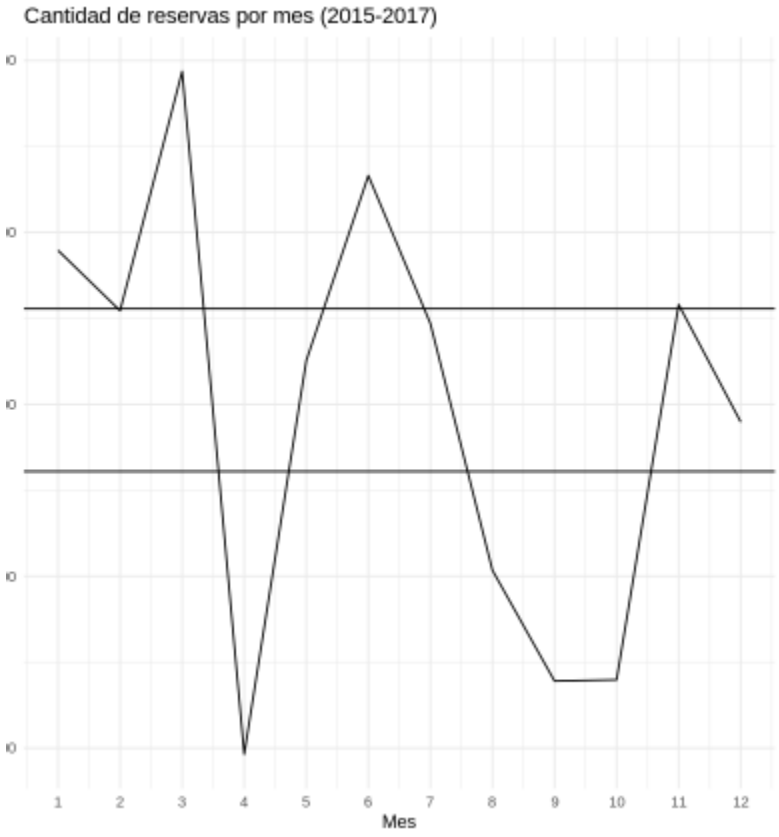
```
[48] df_hotel <- read.csv("hotel_bookings_final.csv")

df_hotel

[50] reservas_hotel <-table(df_hotel$hotel)

print(reservas_hotel)

  City Hotel Resort Hotel
       71397        34605
```

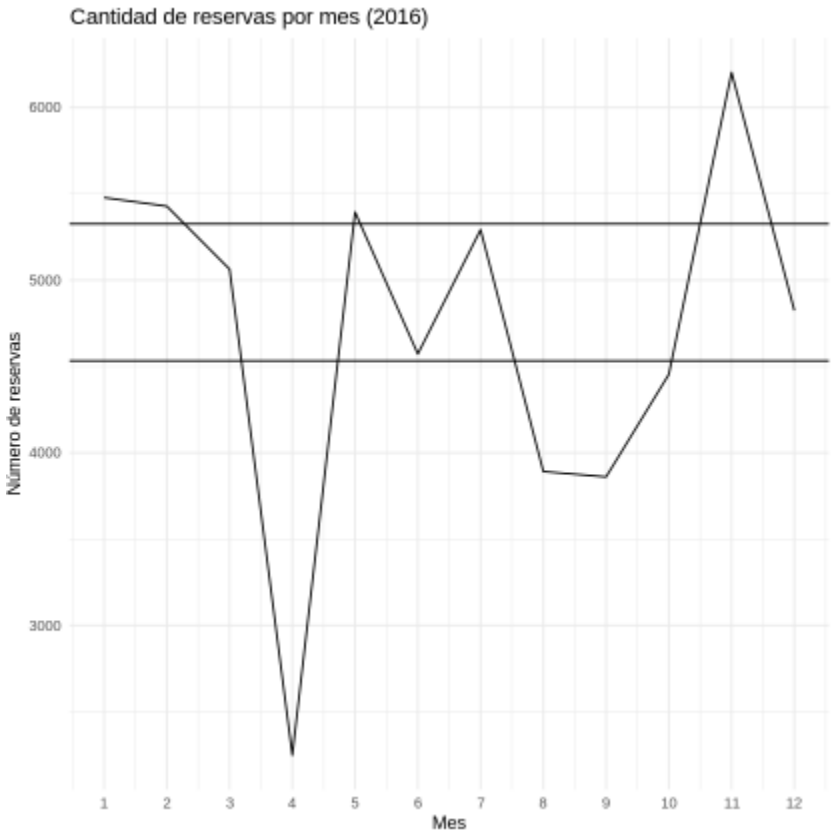
ii. ¿Está aumentando la demanda con el tiempo?

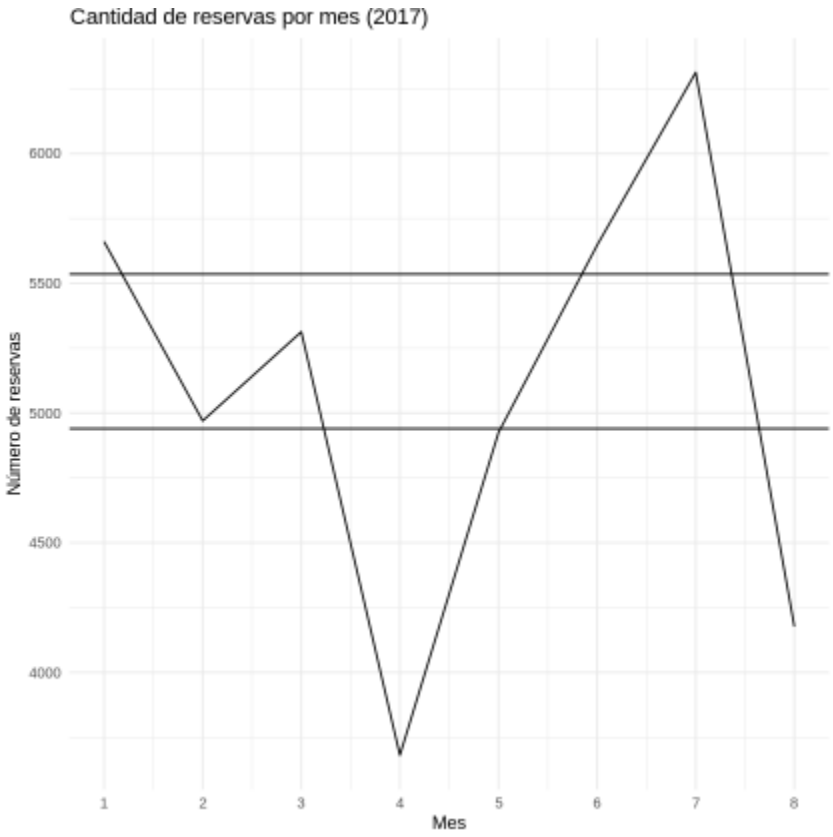


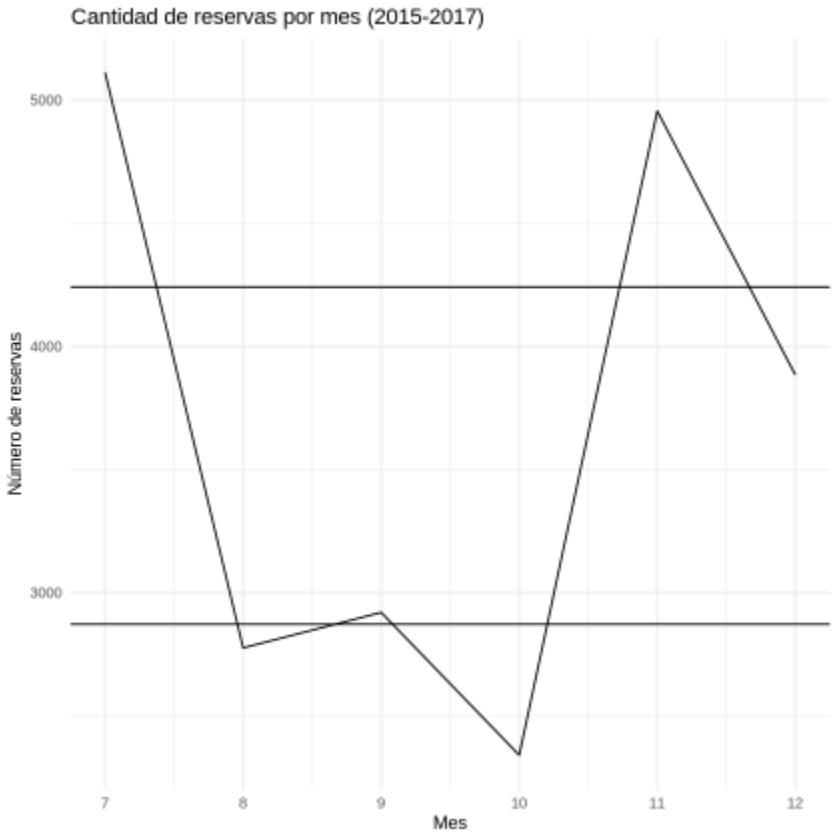
Existe una demanda que aumenta notablemente desde el año 2015 al 2017

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


Cantidad de reservas por mes (2015-2017)
1
2
3
4
5
6
7
8
9
10
11
12
Mes

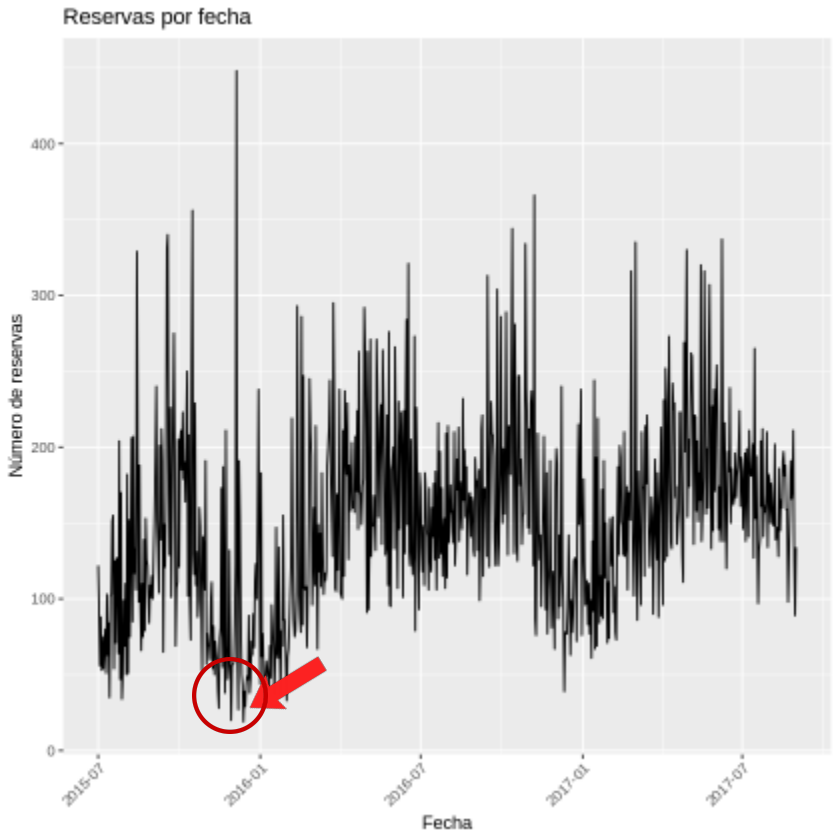

Cantidad de reservas por mes (2016)
Número de reservas
6000
5000
4000
3000
1
2
3
4
5
6
7
8
9
10
11
12
Mes


Cantidad de reservas por mes (2017)
Número de reservas
6000
5500
5000
4500
4000
1
2
3
4
5
6
7
8
Mes



- Como se visualiza en la gráfica, notamos que los meses en los que más reservas se realizan durante los 3 años son en mayo, julio, agosto y octubre, siendo Agosto el mes con más reservas de los tres.
- Los meses con de baja temporada de reservan son enero, febrero, noviembre y diciembre. De estos, Enero es donde la temporada es mas baja.
- La temporada media esta conformada por los meses de marzo, abril, junio y setiembre.

iv. ¿Cuándo es menor la demanda de reservas?



A grouped_df: 6 × 4

arrival_date_year	arrival_date_month	arrival_date_day_of_month	n
<int>	<chr>	<int>	<int>
2015	December	13	19

En este ejercicio usamos un summarise para contar las filas de cada fecha, luego la ordenamos de menor a mayor.

La menor demanda de reservas se da el 13 de diciembre de 2015 con solo 19 reservas

v. ¿Cuántas reservas incluyen niños y/o bebes?

Grafico del numero de reservas con bebés:

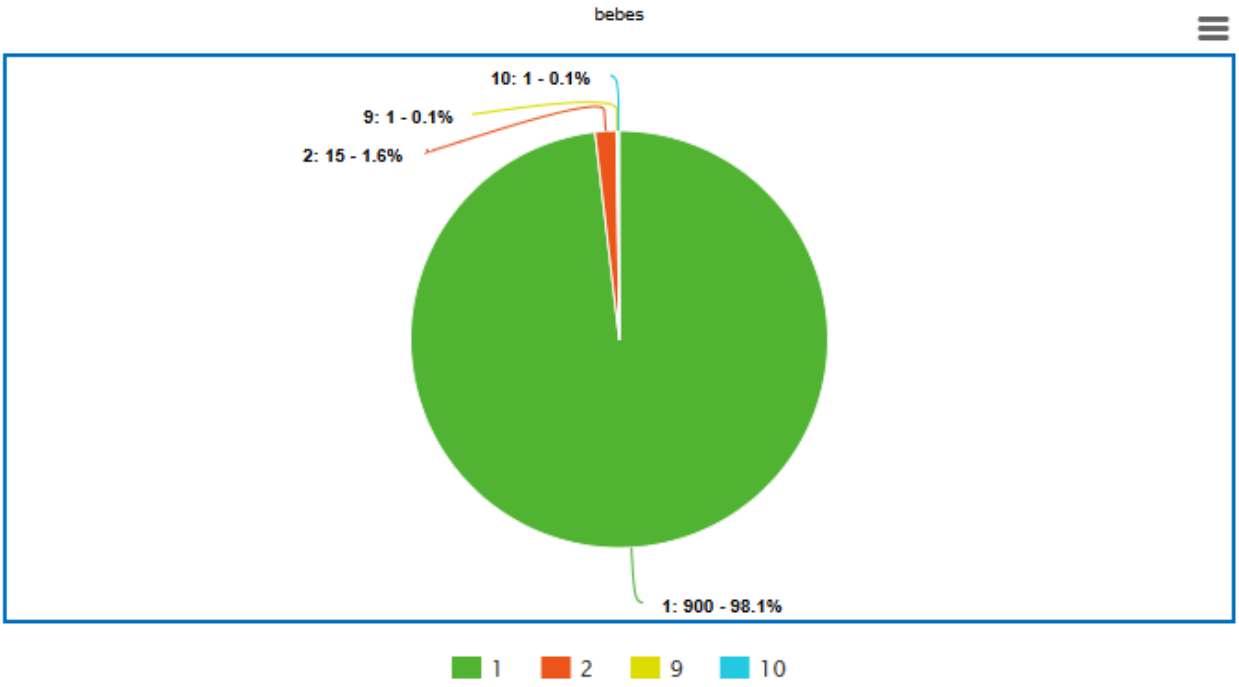
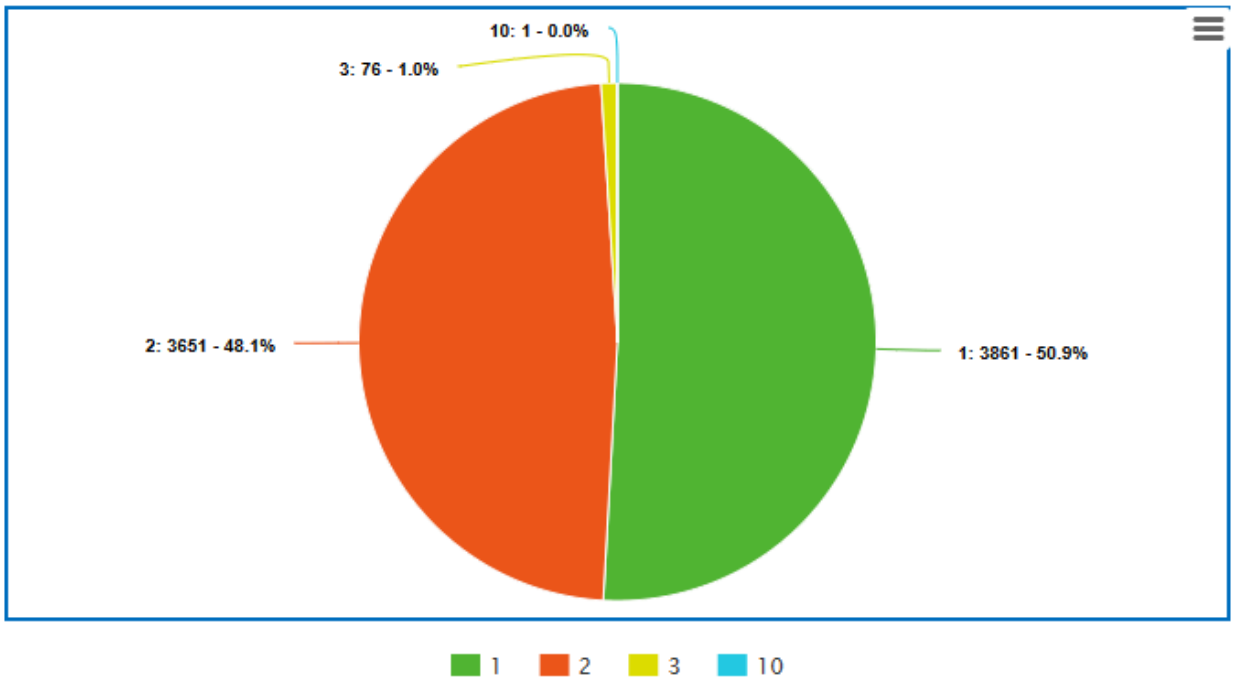


Grafico del numero de reservas con niños:



```
suma_columna <- sum(suma_sin_0)
print(suma_columna)
```

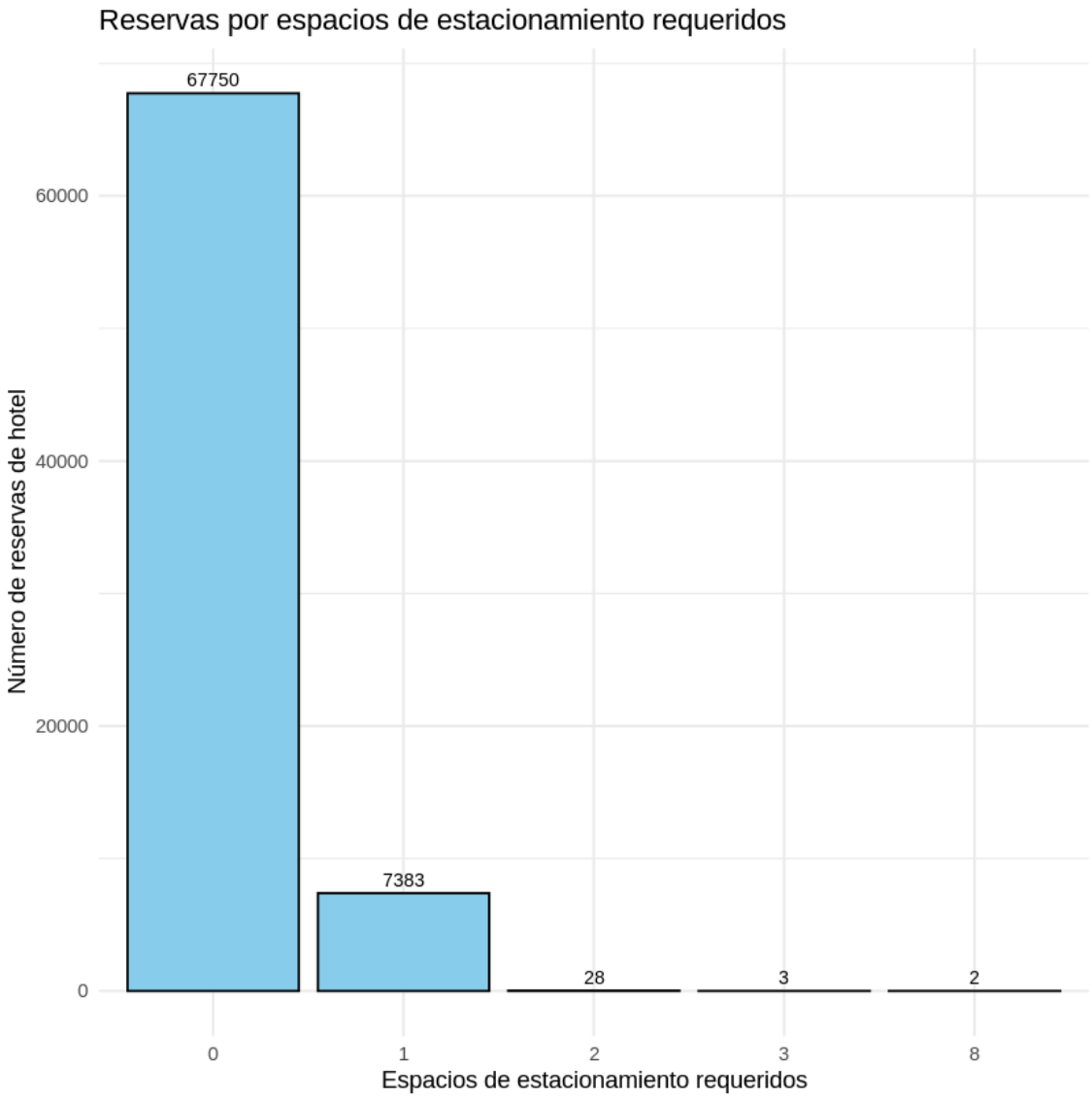
```
[1] 9507
```

En este ejercicio usamos un count en la cantidad de niños y bebés para sumarlos, en la suma quitamos el primer valor del vector porque es los que no tienen ni niños ni bebés. De ahí sumamos toda la columna para ver el total.

La cantidad de reservas que incluyen niños y/o bebés es 6777

vi. ¿Es importante contar con espacios de estacionamiento?

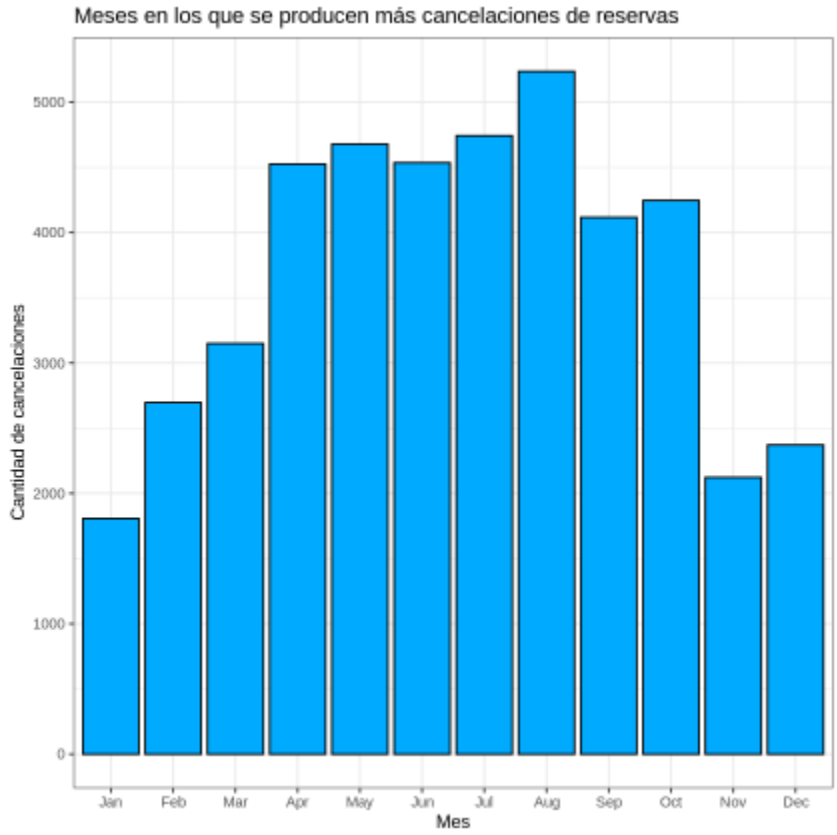
required_car_parking_spaces	n
<int>	<int>
0	67750
1	7383
2	28
3	3
8	2



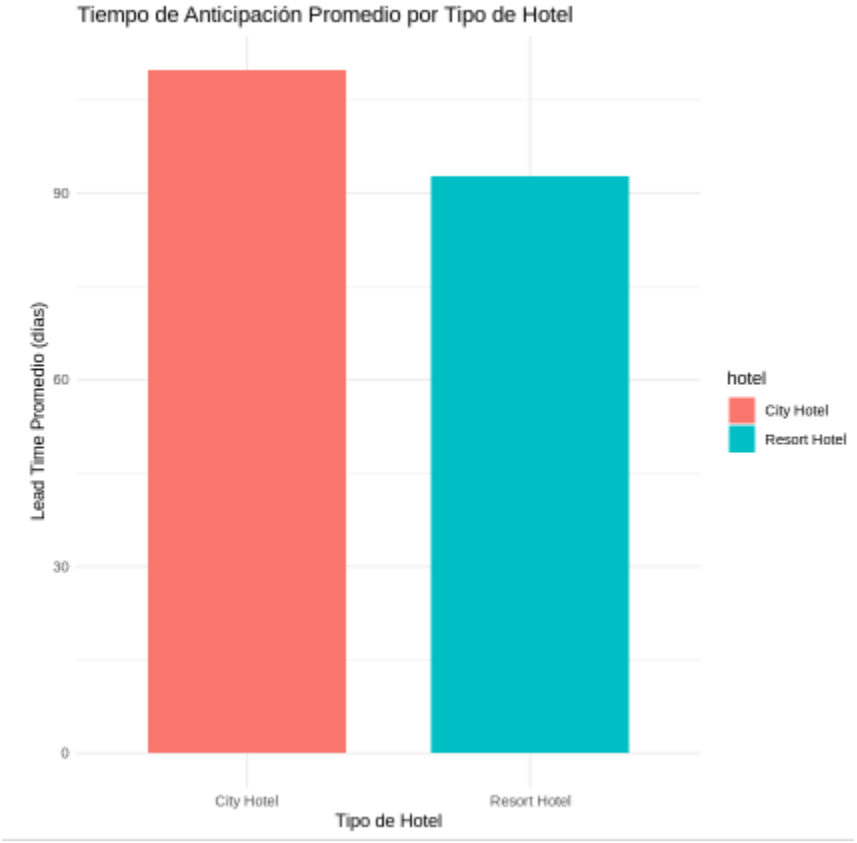
A pesar de que la mayoría de reservas no requieren estacionamientos, es importante porque se puede visualizar que incluso hay reservas que requirieron hasta de 8 estacionamientos.

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

Agosto y julio son los meses en los que ocurrió la mayor cantidad de cancelaciones de reservas



viii. ¿Cuál es el tiempo de anticipación promedio para reservas en diferentes tipos de hotel?

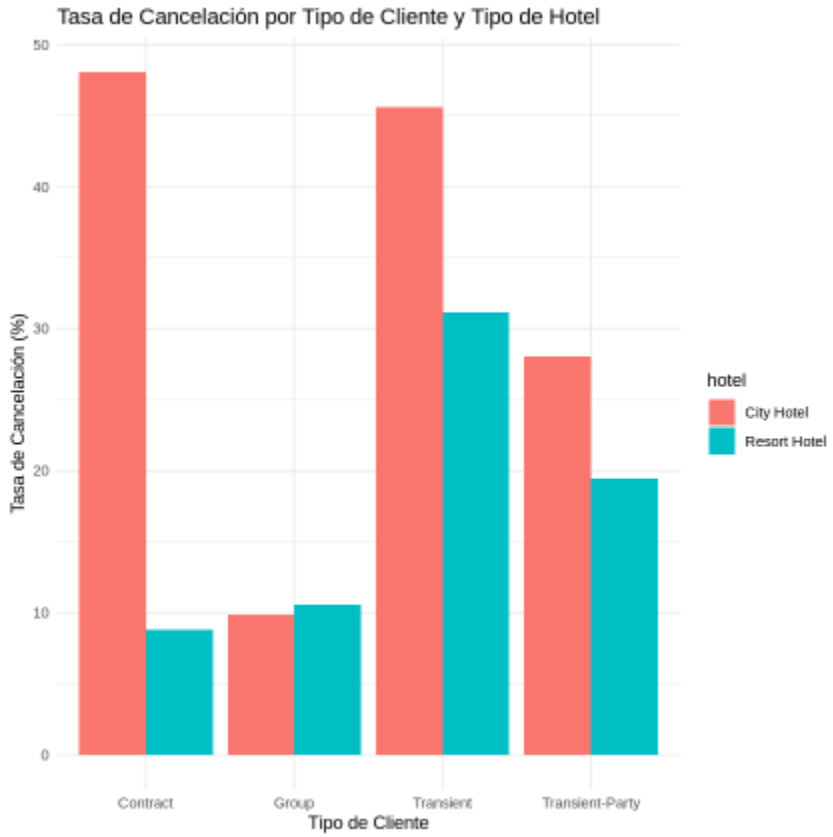


Viendo el análisis mediante el gráfico podemos observar que el hotel resort tiende a reservar con menor anticipación. Esto puede ser que el hotel urbano tiene mejores estrategias para que los clientes quieran reservas antes.

ix.¿Qué porcentaje de reservas son de última hora (por ejemplo, realizadas dentro de una semana antes de la fecha de llegada)?

Este porcentaje nos muestra las reservas hechas en la última semana, el 16.5% de las reservas totales fueron hechas en la última semana.

x. ¿Cuál es la relación entre la tasa de cancelación y el tipo de cliente?



Se puede observar que la tasa de cancelación es mayor para el hotel urbano del grupo de clientes por contrato, y en el caso del hotel resort el grupo que más cancela es el de los transient o temporales.

4. Conclusiones Preliminares

Basado en el análisis de los datos de reservas de hotel, se puede elaborar un perfil detallado del cliente típico. Este perfil nos permite entender mejor las tendencias y preferencias de los clientes:

- **Preferencia del hotel:** Los clientes tienden a preferir los hoteles de ciudad. Lo que puede indicar que los viajes de negocios o las visitas a las áreas urbanas son más comunes que las vacaciones en destinos más rurales o aislados.
- **Temporada de reservas:** La temporada baja ocurre durante los meses de noviembre, diciembre, enero y febrero.
- **Cancelaciones de reservas:** Los meses con más cancelaciones son agosto y julio.

- **Necesidad de estacionamiento:** Se necesitan como mínimo 4 estacionamientos por hotel. Lo que sugiere que muchos clientes optan por alquilar un auto durante su estancia.
- **Familia:** La mayoría de las reservas no incluyen a familias con niños o bebés. Lo que llega a indicar que los viajes en solitario o en pareja son más comunes.

5. Archivar y Publicar

Los archivos subidos a github se encuentran en:
https://github.com/Alecit13/CC216-TP-2024-1.git