Universidad Peruana de Ciencias Aplicadas

Trabajo Parcial

Curso: Data Science

Sección: CC52

Profesor: Nérida Isabel Manrique Tunque

Código del alumno	**Apellidos y Nombres completos**
U202115844	Mancilla Cienfuegos, Paula Jimena
U202110078	Loyola Huaman, Jose Alejandro
U20201E766	Vargas Soto, Lennin Jhair
U20231D424	Caro Leon, Lesly Estefany

Septiembre, 2023

Caso de Análisis

Explicación sobre el origen de los datos (procedencia de los datos, autor/autores, fecha, país, etc.

El origen de los datos proviene de dos conjuntos de datos sobre la demanda de dos hoteles ubicados en Lisboa, Portugal. Uno de los hoteles (H1) es un hotel tipo resort, mientras que el otro (H2) es un hotel en la ciudad.Los datos abarcan reservas con fechas de llegada entre el 1 de julio de 2015 y el 31 de agosto de 2017, incluyendo tanto reservas que llegaron efectivamente como aquellas que fueron canceladas.Estos conjuntos de datos tienen un valor importante para la investigación y la educación en áreas como la gestión de ingresos, el aprendizaje automático o la minería de datos.

Caso de Usos

Casos de uso aplicables (describir, por ejemplo: ¿Para quién sería importante el análisis de estos datos?, ¿Quien o quienes se benefician?

1. Gestión Hotelera

Casos de uso:

Optimización de precios: Analizar la demanda en diferentes temporadas para establecer una estrategia de precios dinámica.

Gestión de inventario: Administrar mejor los recursos del hotel, como habitaciones, estacionamientos, etc.

Personalización del servicio: Entender las preferencias de los clientes para ofrecer servicios personalizados.

Beneficiarios:

Gestores de hoteles: Tomar decisiones basadas en datos para mejorar el rendimiento del negocio.

Empleados: Optimización de las operaciones, lo que podría llevar a un entorno laboral más eficiente.

2. Clientes

Casos de uso:

Recomendaciones personalizadas: Crear sistemas de recomendación que sugieran ofertas y paquetes basados en el perfil del cliente.

Experiencia mejorada: Entender las preferencias y necesidades de diferentes grupos de clientes (familias con niños, parejas, viajeros de negocios, etc.) para mejorar su experiencia durante la estadía.

Beneficiarios:

Clientes: Obtienen una experiencia más personalizada y satisfactoria durante su estancia.

3. Agencias de Viajes y Plataformas de Reservas Online

Casos de uso:

Análisis de mercado: Comprender las tendencias del mercado para desarrollar estrategias de marketing efectivas.

Predicción de demanda: Desarrollar modelos predictivos para anticipar la demanda en diferentes temporadas.

Beneficiarios:

Agencias de viajes: Pueden ofrecer ofertas más atractivas y personalizadas a sus clientes.

Clientes: Obtienen ofertas más personalizadas y pueden planificar mejor sus viajes.

4. Gobierno y Organismos de Turismo

Casos de uso:

Planificación estratégica: Utilizar los datos para planificar estrategias de desarrollo turístico.

Políticas de regulación: Establecer políticas y regulaciones basadas en datos reales para el sector hotelero.

Beneficiarios:

Gobierno: Puede planificar e implementar políticas más informadas.

Sector turístico: Beneficia de políticas bien informadas que pueden ayudar a fomentar el crecimiento del sector.

5. Investigadores y Académicos

Casos de uso:

Investigación de mercado: Realizar estudios académicos y de mercado basados en datos reales del sector hotelero.

Publicaciones académicas: Desarrollar estudios de caso y publicaciones basadas en análisis de datos reales.

Beneficiarios:

Investigadores: Acceden a un rico conjunto de datos para sus estudios.

Estudiantes: Pueden utilizar los datos para proyectos académicos y aprendizaje práctico.

En todos los casos, el EDA ayudará a descubrir insights valiosos, identificar patrones y tendencias, y entender mejor las dinámicas del mercado hotelero, lo que a su vez puede ayudar a tomar decisiones más informadas y desarrollar estrategias más efectivas.

Conjunto de datos (DATA SET)

❖ Descripción de la estructura de los datos (tabla conteniendo la estructura y descripción de cada uno de los datos).

Campo	Descripción	Tipo
hotel	Nombre del hotel	Character
is_canceled	Indica si la reserva fue cancelada (1) o no (0).	Integer
lead_time	Tiempo de espera	Integer
arrival_date_year	Año de llegada	Integer
arrival_date_month	Mes de llegada	Character
arrival_date_week_number	Número de semana de llegada en el año	Integer

arrival_date_day_of_month	Día del mes de llegada	Integer
stays_in_weekend_nights	Número de noches de fin de semana en la estancia	Integer
stays_in_week_nights	Número de noches de entre semana en la estancia	Integer
adults	Número de adultos en la reerva	Integer
children	Número de niños en la reserva	Integer
babies	Número de bebés en la reserva	Integer
meal	Tipo de comida incluida en la reserva	Character
country	País de origen del huésped	Character
market_segment	Segmento de mercado de la reserva	Character
distribution_channel	Canal de distribución de la reserva	Character
is_repeated_guest	Indica si el huésped es repetido (1) o no (0)	Integer
previous_cancellations	Número de reservas canceladas previamente	Integer
previous_bookings_not_canceled	Número de reservas no canceladas previamente	Integer
reserved_room_type	Tipo de habitación reservada	Character
assigned_room_type	Tipo de habitación asignada	Character
booking_changes	Número de cambios realizados en la reserva	Integer
deposit_type	Tipo de depósito utilizado para la reserva	Character

agen	Agente de viajes que realizó la reserva	Character
company	Empresa asociada a la reserva	Character
days_in_waiting_list	Número de días en lista de espera	Integer
customer_type	Tipo de cliente (Transient, Contract, Group, etc.).	Character
adr	Tarifa diaria promedio	Numeric
required_car_parking_spaces	Número de espacios de estacionamiento requeridos	Integer
total_of_special_requests	Número total de solicitudes especiales	Integer
reservation_status	Estado de la reserva (Check-Out, Canceled, etc.)	Character
reservation_status_date	Fecha del estado de la reserva	Character

Análisis Exploratorio de datos

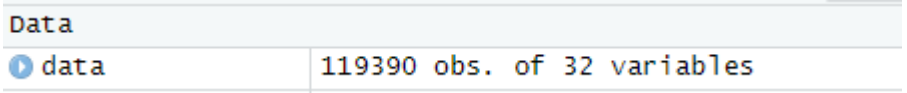
Descripción de instrucciones ejecutadas en R/RStudio y resultados obtenidos para:

❖ CARGAR DATOS

Se realiza la carga del dataset donde consideramos los parámetros header = TRUE, stringsAsFactors = FALSE

```
# Cargamos los datos
data <- read.csv("C:/Users/pjman/Downloads/hotel_bookings.csv", header = TRUE, stringsAsFactors = FALSE)
```

Así mismo, se verifica si la carga de datos fue realizada de manera exitosa.



❖ INSPECCIONAR DATOS

En nuestra base de datos contamos con 3 distintos tipos de datos al indagar sobre el contenido, obtenemos los distintos tipos de los mismos:

Integer: arrival_date_week_number, arrival_date_day_of_month, stays_in_weekend_nights, stays_in_week_nights, adults, children, babies, is_canceled, lead_time, arrival_date_year.

Estas variables son de tipo entero y generalmente se utilizan para representar valores numéricos enteros

Character: hotel, arrival_date_month, meal, country, market_segment, distribution_channel, reserved_room_type, assigned_room_type, deposit_type, agen, company, customer_type, reservation_status, reservation_status_date.

Estas variables son de tipo carácter y contienen texto o cadenas de caracteres.

Numeric: adr.
La variable "adr" es de tipo numérico que se se utiliza para representar valores numéricos continuos, como tarifas diarias promedio.

❖ PRE-PROCESAR DATOS

Identificación de datos faltantes:

Se identifican los datos faltantes en el dataset con los siguientes comandos que se encargan de contar los datos faltantes de cada columna.

```
# Indentificación de datos faltantes en general
# Contamos los valores faltantes en cada columna del conjunto de datos 'data'

missing_values <- colSums(is.na(data))

# Mostramos la cantidad de valores faltantes por columna

print(missing_values)
```

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

En el siguiente cuadro se ven la cantidad de valores faltantes por columna. En este caso solo hay valores faltantes en la columna “children”, en donde el código ha contado 4 datos faltantes.

Explicación y aplicación de la técnica utilizada para eliminar o completar los datos faltantes

Una vez identificados se procede a eliminar las filas de los datos faltantes

```
# Eliminamos filas con valores NA y almacenar el resultado en un nuevo conjunto de datos

data_sin_na <- na.omit(data)
```

Con este comando se crea una nueva cvs llamado data_sin_na donde se alojarán nuestra data omitiendo las filas donde no haya información.

Otra alternativa es filtrar las filas sin valor lo cual se hace con el siguiente comando

```
# Otra alternativa es:
# Filtrar las filas completas (sin valores faltantes) en el conjunto de datos

data_filtras_filas_completas <- data[complete.cases(data), ]
```

Complete.cases(data): crea un vector lógico que indica qué filas de data están completas y cuáles no.

data[complete.cases(data),]: solo se seleccionan las filas para las cuales complete.cases(data) es TRUE.

Por lo que el último comando es más directo y conveniente para obtener los mismos resultados. Al hacer el filtrado o eliminación de datos estamos eliminando los datos inconsistentes, que se puede dar por un error en el registro ya sea manual o informático.

data	119390 obs. of 32 variables
data_filtras_fi…	119386 obs. of 32 variables
data_sin_na	119386 obs. of 32 variables

La cantidad de resultados obtenidos en ambas tablas tanto de data_filtras_fi como de data_sin_na seran los mismos. Otra solución ante el problema, podría ser aplicar la media en la columna de los datos faltantes y completarlos. Pero

Identificación de datos atípicos (Outliers)

Explicación y aplicación de la(s) técnica(s) utilizada(s) para transformar los datos atípicos.

La identificación y aplicación de los outliers es una parte crucial para el análisis de datos, ya que son los que tendrán un impacto directo con los resultados del análisis estadístico así como en la creación de modelos predictivos.

Primero se identifican los datos atípicos esto lo hacemos con el siguiente código

```r
# Creamos una lista vacía para almacenar las tablas de frecuencia

frequency_tables <- list()

# Iteramos a través de cada columna del conjunto de datos

for (col_name in colnames(data_sin_na)) {
  frequency_tables[[col_name]] <- table(data_sin_na[[col_name]])
}

# Mostramos las tablas de frecuencia para cada columna

for (col_name in colnames(data_sin_na)) {
  cat("Tabla de frecuencia para la columna:", col_name, "\n")
  print(frequency_tables[[col_name]])
}

# Identificamos los valores atipicos
# Identificamos outliers en las columnas numéricas y enteras
# Seleccionamos todas las columnas numéricas y enteras

numeric_integer_columns <- data_sin_na[sapply(data_sin_na, function(x) is.numeric(x) || is.integer(x))]

# Creamos una lista para almacenar los resultados de outliers

outliers_results <- list()
```

❖ VISUALIZAR DATOS

Con el fin de no perder información y abarcar todo los campos posibles, se optó por utilizar todos los datos brindados.

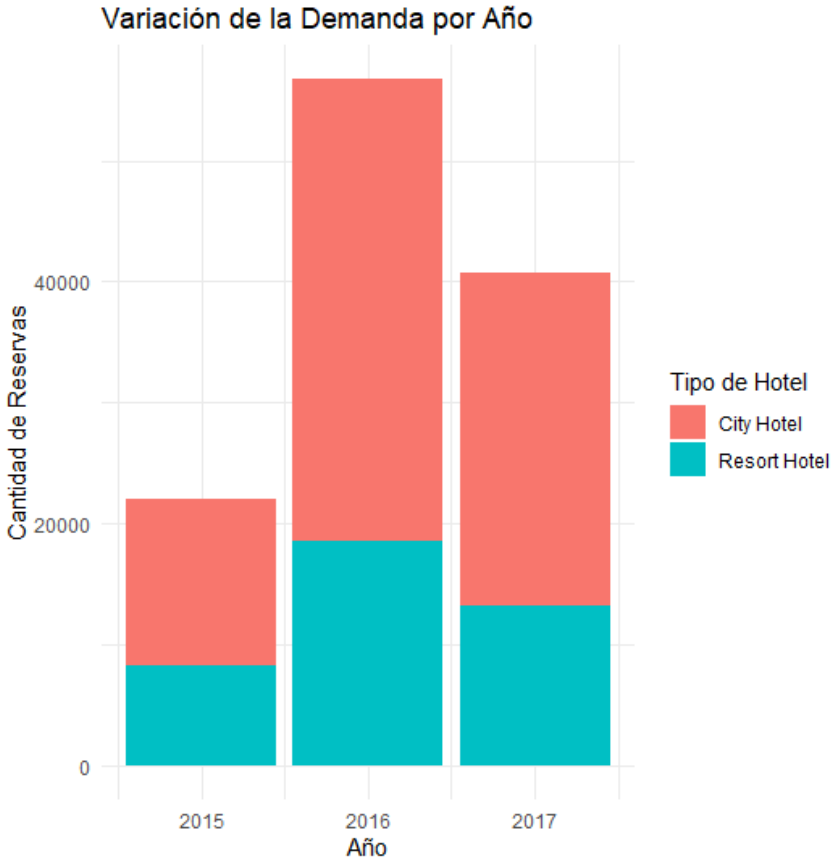
a. ¿Cuántas reservas se realizan por tipo de hotel? o ¿Qué tipo de hotel prefiere la gente?

Como se muestra en la gráfica mostrada se puede ver que City Hotel cuenta con 80000 reservas y Resort Hotel con 40000

gráfica 1

b. ¿Está aumentando la demanda con el tiempo?

Sí, progresivamente en aumento.



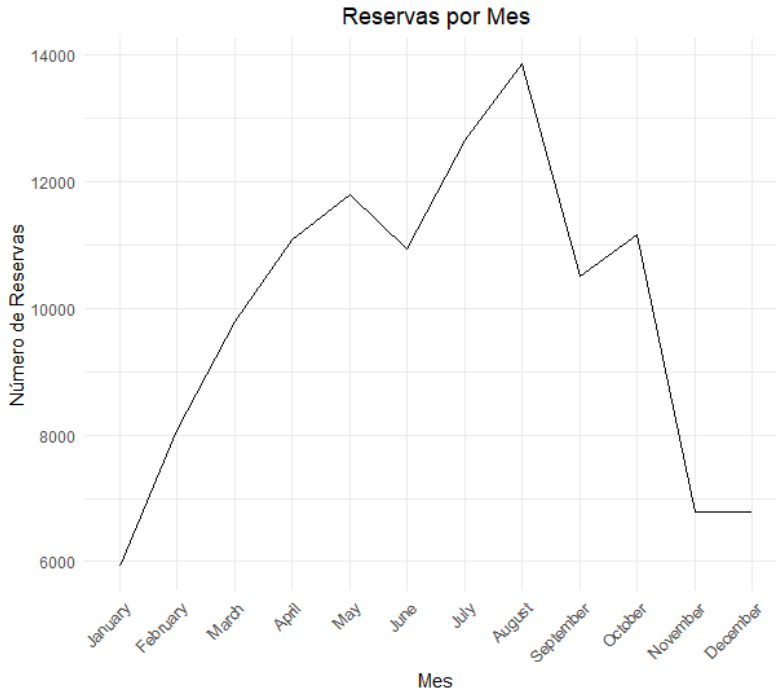
gráfica 2

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

En Enero se puede observar que el número de reservas es menor: 6 000

En Marzo y Octubre se puede observar que el número de reservas es medio: 8 000

En Agosto se puede observar que el número de reservas es mayor: 14 000



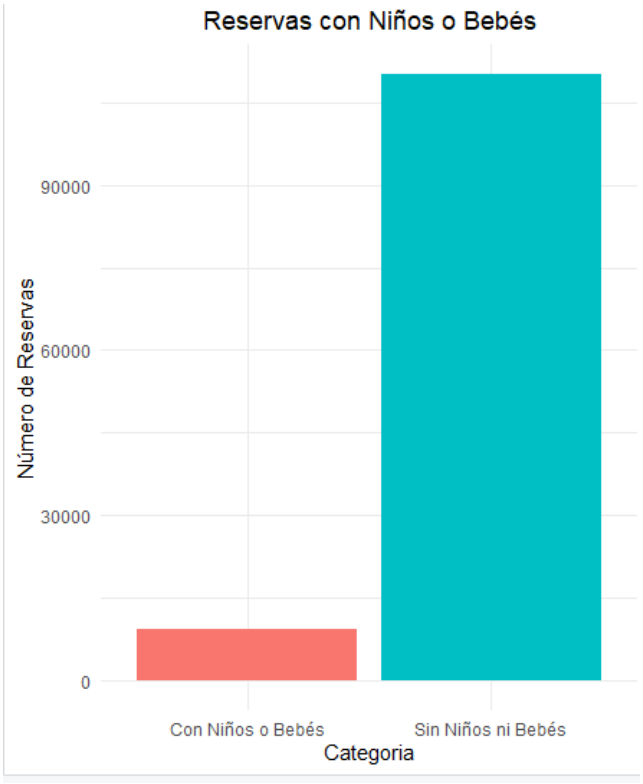
gráfica 3

d. ¿Cuándo es menor la demanda de reservas?

La menor cantidad de reservas según la gráfica mostrada es Enero: 6 000

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

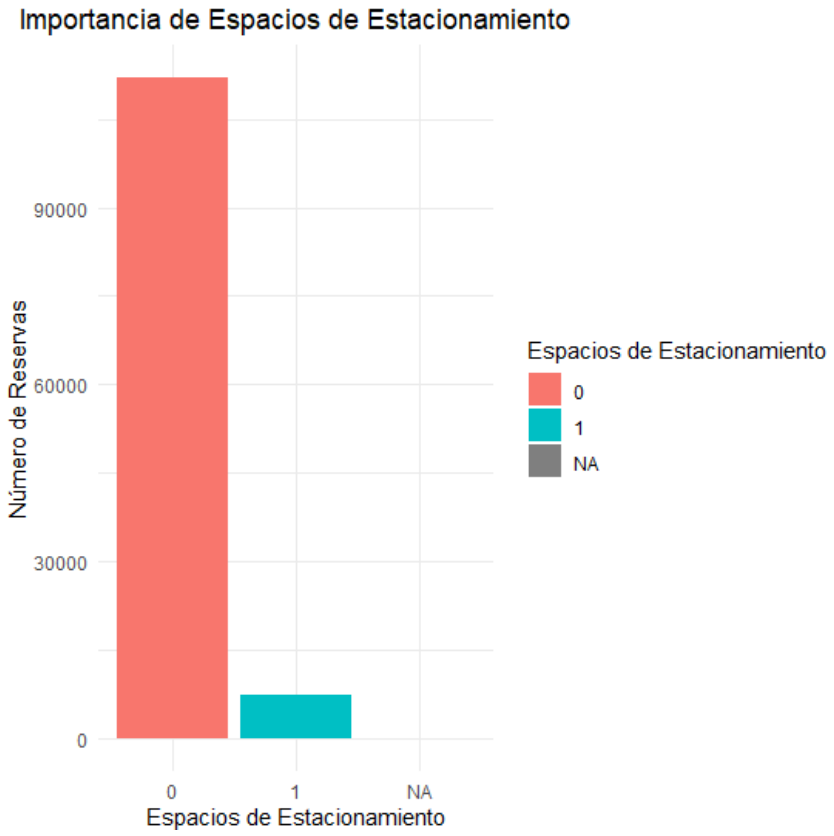
Reservas que incluyen niños y/o bebés: 1500



gráfica 4

f. ¿Es importante contar con espacios de estacionamiento?

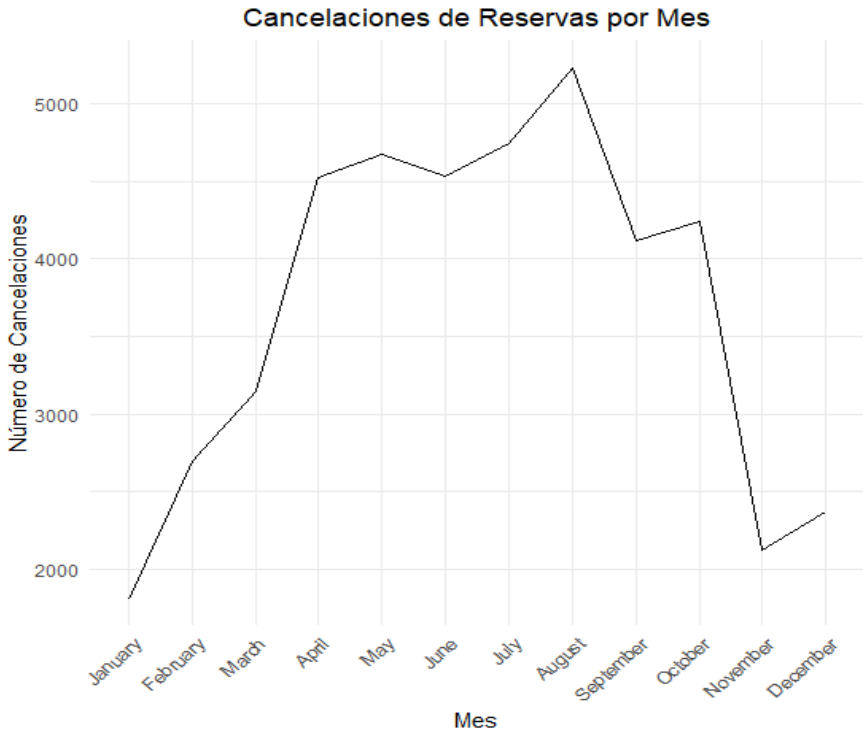
No se considera importante priorizar un espacio de estacionamiento, ya que en casi todos los casos los huéspedes no llegan con un auto propio.



gráfica 5

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

Se producen en Agosto: 5500 cancelaciones



gráfica 6

Conclusiones Preliminares

1. Analizando la gráfica 1 de la pregunta a. Se puede concluir que, la diferencia entre la cantidad de reservas entre uno y otro es abismal, aproximadamente 40 000 reservas de diferencia. Este resultado se puede deber al tamaño de cada hotel, sus distintas estrategias de marketing, variables externas a las variables que tenemos.

2. Analizando la gráfica 2 de la pregunta b. Se puede concluir que, existe un aumento significativo de demanda desde el 2015 al 2017. Con esta información, se pueden estimar el número de reservas que se harán el siguiente año y planificar los gastos. Sin embargo, hay que tener en cuenta factores externos que no estamos considerando, en las variables que también contribuyeron al crecimiento de las reservas en esos años.

3. Analizando la gráfica 3 de la pregunta c. Se puede concluir que de Noviembre a Enero, el número de reservas baja, por lo que los precios, también podrían bajar para motivar a los potenciales clientes a quedarse en el hotel. También se observa, como de Marzo a Octubre hay una concurrencia media, llegando hasta 80000 reservas. Por último se observa que en Agosto es el mes con más reservas, por lo que aprovechando la temporada se puede prestar a subir los precios habilitar áreas del hotel, etc.

4. Analizando la gráfica 3 de la pregunta d. Se puede concluir que la menor cantidad de reservas se encuentra en Enero llegando a menos de 6000 reservas en ese mes.

5. Analizando la gráfica 4 de la pregunta e. Se puede concluir que las reservas con niños es mínima llegando a un máximo de 15000 reservas con niños contra 100 000 reservas sin niños. Por lo que personalizar áreas para niños no sería una prioridad.

6. Analizando la gráfica 5 de la pregunta f. Se puede concluir que en este tipo de hoteles no es importante priorizar el estacionamiento como en el caso de un hotel que esté en carretera que está creado exclusivamente como punto de paso y su público principal son personas que estén viajando y necesiten un breve descanso.

7. Analizando la gráfica 6 de la pregunta g. Se puede concluir que como Agosto es el mes de más reservas y hay más concurrencia, también es el mes donde hay más errores, por ello se tiene que ver por eso se tiene que prever al momento de estimar cuánto flujo hay en cada més no basta con considerar las reservas sino también las cancelaciones para ver el flujo real del hotel.