Universidad Peruana de Ciencias Aplicadas

Facultad de Ingeniería
TB1

Ciclo 05

Nombre del curso
Fundamentos de Data Science

Sección:
258

Nombre del profesor
Nerida Isabel Manrique Tunque

"Informe de TB1"

Carrera:
Ciencias de la computación

Nombre del caso:
Hotel booking

Relación de integrantes:

Alzamora Gonzales Leonel - U20231c427
Avalos Sánchez César Gabriel - U202310307
Rivas Pinto, Piero Aldair - U202122405
Rojas Cuadros, Fabian Marcelo - U202218498

Mes y año:
Mayo 2025

1. **CASO DE ANÁLISIS**
    a. **Origen de los Datos:**
    El conjunto de datos “Hotel Booking Demand” fue originalmente recopilado por los investigadores Nuno Antonio, Ana de Almeida y Luis Nunes como parte de un estudio publicado en la revista Data in Brief (2018) bajo el título Hotel booking demand datasets https://www.sciencedirect.com/science/article/pii/S2352340918315191.

    Estos datos provienen de registros administrativos de dos hoteles ubicados en Portugal: uno de tipo urbano (City Hotel) y otro de tipo resort (Resort Hotel). Se recolectaron entre julio de 2015 y agosto de 2017. La fuente original, alojada en Kaggle, fue modificada para fines educativos en esta evaluación, introduciendo valores faltantes y outliers deliberadamente para ejercitar técnicas de limpieza y análisis de datos. La base es considerada confiable, ya que fue construida a partir de datos reales y revisada por pares para su publicación científica.

    b. **Casos de uso aplicable:**
        i. Este análisis puede ser de gran utilidad para:
        - Gerentes y administradores de hoteles (para mejorar la gestión de reservas y la ocupación).

        - Departamentos de marketing (para planificar campañas en temporadas altas o bajas).

        - Agencias de viajes (para conocer tendencias de demanda).

        - Empresas de análisis turístico o consultoras en hotelería.

        ii. ¿Qué problemas o necesidades responde este análisis? (ejemplo, optimización de ocupación, predicción de demanda, etc)
        - **Optimización de ocupación hotelera:** identificar períodos de alta demanda para ajustar precios y personal.
        - **Predicción de cancelaciones:** comprender en qué momentos o bajo qué condiciones se producen más cancelaciones.
        - **Segmentación de clientes:** distinguir perfiles de huéspedes según tipo de reserva, duración de estancia o características familiares.
        - **Toma de decisiones estratégicas:** fundamentar decisiones de inversión o expansión a partir del comportamiento histórico de las reservas.

2. **CONJUNTO DE DATOS (DATA SET)**
    a. **Descripción del Data Set:**

Proporcionar una tabla que describa las variables contenidas en el conjunto de datos, especificando: Nombre de la variable, Tipo de dato (numérico, categórico, etc), descripción breve del significado de cada variable.

Nombre de variable	Tipo de dato		Descripción
hotel	Categórico	Nominal	Tipo de hotel: ciudad o resort.
is_canceled	Booleano		Indica si la reserva fue cancelada (0: no, 1: sí).
lead_time	Numérico	Continuo	Días entre la reserva y la llegada.
arrival_date_year	Numérico	Discreto	Año de llegada.
arrival_date_month	Categórico	Ordinal	Mes de llegada.
arrival_date_week_number	Numérico	Discreto	Número de semana del año en que llega el huésped.
arrival_date_day_of_month	Numérico	Discreto	Día del mes en que llega el huésped.
stays_in_weekend_nights	Numérico	Discreto	Noches de fin de semana de la estadía.
stays_in_week_nigths	Numérico	Discreto	Noches entre semana de la estadía.
adults	Numérico	Discreto	Número de adultos.
children	Numérico	Discreto	Número de niños.
babies	Numérico	Discreto	Número de bebés.
meal	Categórico	Nominal	Tipo de comida reservado.
country	Categórico	Nominal	País de origen del cliente (código ISO).
market_segment	Categórico	Nominal	Segmento de mercado que generó la reserva.
distribution_channel	Categórico	Nominal	Canal a través del cual se realizó la reserva.
is_repeated_guest	Booleano		Indica si el huésped es repetido (1) o no (0).

previous_cancellations	Numérico	Discreto	Número de cancelaciones anteriores.
previous_bookings_not_canceled	Numérico	Discreto	Número de reservas previas no canceladas.
reserved_room_type	Categórico	Nominal	Tipo de habitación reservada.
assigned_room_type	Categórico	Nominal	Tipo de habitación asignada.
booking_changes	Numérico	Discreto	Número de cambios realizados a la reserva.
deposit_type	Categórico	Nominal	Tipo de depósito: ninguno, reembolsable o no reembolsable.
agent	Categórico	Nominal	ID del agente que gestionó la reserva (o NULL).
company	Categórico	Nominal	ID de la empresa asociada (si aplica).
days_in_waiting_list	Numérico	Discreto	Días en lista de espera.
customer_type	Categórico	Nominal	Tipo de cliente (transitorio, contrato, grupo).
adr	Numérico	Continuo	Tarifa diaria promedio (Average Daily Rate).
required_car_parking_spaces	Numérico	Discreto	Espacios de estacionamiento requeridos.
total_of_special_requests	Numérico	Discreto	Número total de solicitudes especiales.
reservation_status	Categórico	Nominal	Estado final de la reserva.
reservation_status_date	Fecha		Fecha en que se registró el estado de la reserva.

3. ANÁLISIS EXPLORATORIO DE DATOS (EDA)

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

a. Cargar datos:

#limpieza

```
rm(list=ls(all=TRUE))
graphics.off()
cat("\014")

#instalamos las librerias
install.packages("ggplot2")
install.packages("dplyr")
install.packages("readr")
install.packages("lubridate")

#cargamos las librerias
library(ggplot2)
library(dplyr)
library(readr)
library(lubridate)

#cargamos el archivo
data_hotel <- read.csv("hotel_bookings.csv", header = TRUE, stringsAsFactors = FALSE)
View(data_hotel)
```

b. Inspeccionar datos:

```
#exploracion general
head(data_hotel)
str(data_hotel)
summary(data_hotel)
names(data_hotel)
dim(data_hotel)
glimpse(data_hotel)

#convertir "NULL" Y "" A NA
data_hotel_modificada <- data.frame(lapply(data_hotel, function(x) {
  x <- as.character(x)
  x[x == "NULL" | x == ""] <- NA
  return(x)
}), stringsAsFactors = FALSE)

#deteccion de duplicados
sum(duplicated(data_hotel_modificada))              # Cuántos duplicados
data_hotel_modificada <- data_hotel_modificada[!duplicated(data_hotel_modificada), ]

#identificamos tipo de variables
num_cols <- names(Filter(is.numeric, data_hotel_modificada))
cat_cols <- names(Filter(is.character, data_hotel_modificada))
```

```
  # Variables numericas
  print(num_cols)
  # Variables categoricas
  print(cat_cols)

  #conversion a factores de variables categoricas
  data_hotel_modificada <- data_hotel_modificada %>%
    mutate(
      hotel = as.factor(hotel),
      meal = as.factor(meal),
      market_segment = as.factor(market_segment),
      distribution_channel = as.factor(distribution_channel),
      reserved_room_type = as.factor(reserved_room_type),
      assigned_room_type = as.factor(assigned_room_type),
      deposit_type = as.factor(deposit_type),
      customer_type = as.factor(customer_type)
    )

  #creacion de variable fecha
  data_hotel_modificada <- data_hotel_modificada %>%
    mutate(arrival_date = dmy(paste(arrival_date_day_of_month, arrival_date_month,
  arrival_date_year)))
```

c. Pre-procesar datos

```
#resumir estadisticas basicas
summary(data_hotel_modificada)
range(as.numeric(data_hotel_modificada$lead_time))
table(data_hotel_modificada$hotel)
table(data_hotel_modificada$meal)
table(data_hotel_modificada$reserved_room_type)

#=================================================================
#identificacion de datos faltantes
#mostramos todos los elementos NA por columna del dataframe original
colSums(is.na(data_hotel_modificada))

#Porcentaje de datos vacios en la columna children
data_hotel_modificar_Children<-data_hotel_modificada
sum(is.na(data_hotel_modificar_Children$children))
mean(is.na(data_hotel_modificar_Children$children)) * 100

#=================================================================
#tratamiento de datos faltantes
```

```
#eliminamos los 4 registros que tienen NA en children porque son una cantidad irrelevante
data_hotel_limpio <-
data_hotel_modificar_Children[!is.na(data_hotel_modificar_Children$children), ]
#verificamos cuantos registros se eliminaron
nrow(data_hotel_modificar_Children) - nrow(data_hotel_limpio)

#porcentaje de datos vacios en la columna company
data_hotel_modificar_Company<-data_hotel_modificada
sum(is.na(data_hotel_modificar_Company$company))
mean(is.na(data_hotel_modificar_Company$company)) * 100

#eliminamos la columna company debido al alto porcentaje de elementos NULL
data_hotel_limpio <- data_hotel_modificar_Company %>% select(-company)


#porcentaje de datos vacios en la columna agent
data_hotel_modificar_Agent<-data_hotel_limpio
sum(is.na(data_hotel_modificar_Agent$agent))
mean(is.na(data_hotel_modificar_Agent$agent)) * 100

#como el porcentaje de elementos vacios en la columna agent no es lo suficientemente
#pequeño como para eliminar los registros ni lo suficientemente grande
#como para eliminar la columna, vamos a rellenar los datos con la moda

#Calculamos la moda
moda <- names(sort(table(data_hotel_limpio$agent), decreasing = TRUE))[1]

#Reemplazamos NA con la moda
data_hotel_limpio$agent[is.na(data_hotel_limpio$agent)] <- moda


#porcentaje de datos vacios en la columna country
data_hotel_modificar_Country<-data_hotel_limpio
sum(is.na(data_hotel_modificar_Country$country))
mean(is.na(data_hotel_modificar_Country$country)) * 100

#como el porcentaje es menor a 1% vamos a eliminar los registros
data_hotel_limpio <-
data_hotel_modificar_Country[!is.na(data_hotel_modificar_Country$country), ]
#verificamos cuantos registros se eliminaron
nrow(data_hotel_modificar_Country) - nrow(data_hotel_limpio)

#=============================================================
#Detectar outliers
```

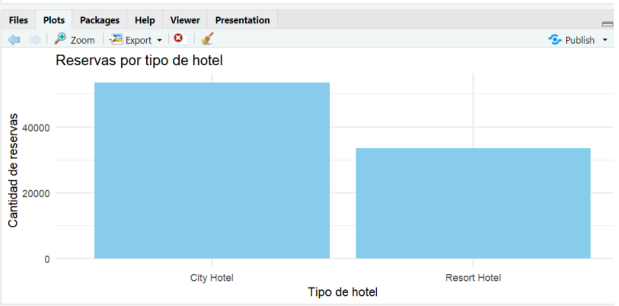
```
#utilizamos diagramas de caja (boxplot) para detectar valores atípicos en lead_time y adr
boxplot(as.numeric(data_hotel_limpio$lead_time), main = "Boxplot de lead_time")
boxplot(as.numeric(data_hotel_limpio$adr), main = "Boxplot de adr")

#=================================================================
#tratamiento de outliers
# Creamos una función para aplicar winsorización (recorta los valores extremos al percentil 1% y 99%)
winsorizar <- function(x, low = 0.01, high = 0.99) {
  x <- as.numeric(x)
  q <- quantile(x, probs = c(low, high), na.rm = TRUE)
  x[x < q[1]] <- q[1]
  x[x > q[2]] <- q[2]
  return(x)
}

#aplicamos la winsorizacion a las columnas lead_time y adr
data_hotel_limpio$adr <- winsorizar(data_hotel_limpio$adr)
data_hotel_limpio$lead_time <- winsorizar(data_hotel_limpio$lead_time)

#revisamos los resultados después del tratamiento de outliers
summary(data_hotel_limpio$adr)
summary(data_hotel_limpio$lead_time)
#guardamos el archivo para el siguiente paso
write.csv(data_hotel_limpio, "hotel_bookings_limpio.csv", row.names = FALSE)

#liberamos los data frame temporales
rm(data_hotel_modificar_Children)
rm(data_hotel_modificar_Agent)
rm(data_hotel_modificar_Country)
rm(data_hotel_modificar_Company)
```

d. Visualización de datos

- ¿Cuántas reservas se realizan por tipo de hotel? ¿Qué tipo de hotel prefiere la gente?
    - Se realizaron 53422 reservas en el “City Hotel” y 33522 en el “Resort Hotel”, la gente prefiere el “City Hotel”.

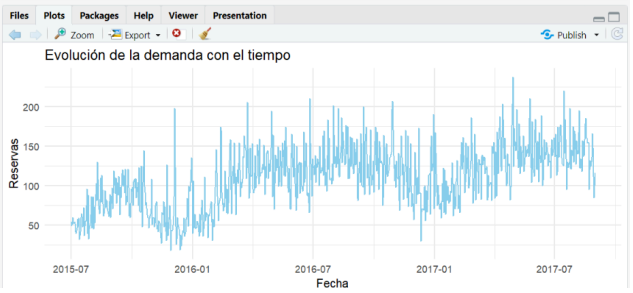
```r
#1) ¿Cuantas reservas se realizaron por tipo de hotel?¿Que tipo de hotel prefiere la gente?
table(data_hotel_limpio$hotel)
prop.table(table(data_hotel_limpio$hotel)) * 100

ggplot(data_hotel_limpio, aes(x = hotel)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Reservas por tipo de hotel", x = "Tipo de hotel", y = "Cantidad de reservas") +
  theme_minimal()
```

```
City Hotel Resort Hotel
      53422        33522
City Hotel Resort Hotel
   61.44415     38.55585
```



- ¿Está aumentando la demanda con el tiempo?
  - La demanda efectivamente está aumentando con el paso del tiempo.

```r
#2) ¿Está aumentando la demanda con el tiempo?
data_hotel_limpio %>%
  group_by(arrival_date) %>%
  summarise(reservas = n()) %>%
  ggplot(aes(x = arrival_date, y = reservas)) +
  geom_line(color = "skyblue") +
  labs(title = "Evolución de la demanda con el tiempo", x = "Fecha", y = "Reservas") +
  theme_minimal()
```



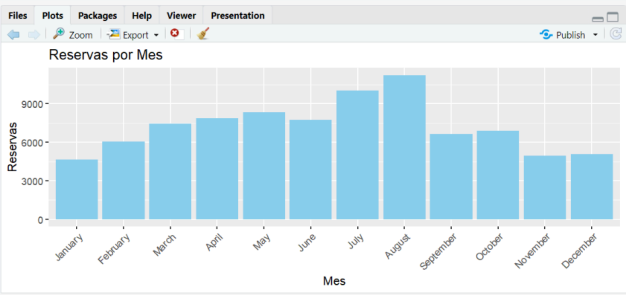
- ¿Cuáles son las temporadas de reservas (alta, media, baja)?
  - Podemos observar que en la temporada de verano es donde aumenta(June ,July, August). En la temporada de primavera(Otoño) es media(April, May), En la temporada de Invierno es baja(November y December).

```r
#3) ¿Cuáles son las temporadas de reservas (alta, media, baja)?
data_hotel_limpio %>%
  group_by(arrival_date_month) %>%
  summarise(reservas = n()) %>%
  arrange(desc(reservas))

data_hotel_limpio$arrival_date_month <- factor(data_hotel_limpio$arrival_date_month,
                                               levels = month.name)

ggplot(data_hotel_limpio, aes(x = arrival_date_month)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Reservas por Mes", x = "Mes", y = "Reservas") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
  arrival_date_month reservas
  <fct>              <int>
1 August             11236
2 July               10024
3 May                 8343
4 April               7871
5 June                7753
6 March               7459
7 October             6883
8 September           6659
9 February            6043
10 December           5082
11 November           4953
12 January            4638
```

```r
data_hotel_limpio$arrival_date_month <- factor(data_hotel_limpio$arrival_date_month,
```

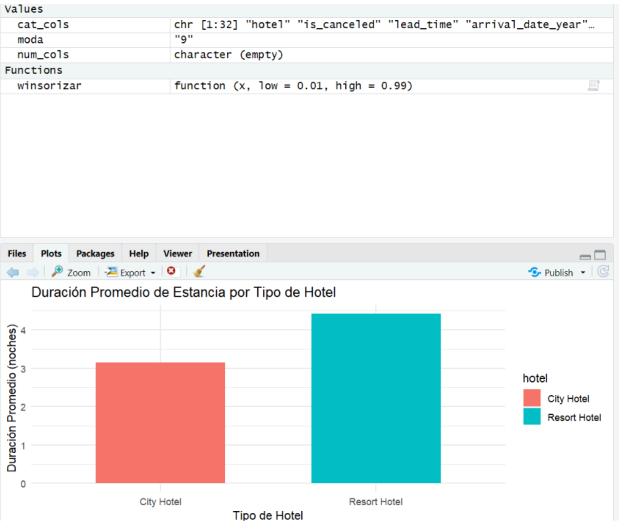
```
Functions
  winsorizar    function (x, low = 0.01, high = 0.99)
```



- ¿Cuál es la duración promedio de las estancias por tipo de hotel?
  - La duración promedio en el City Hotel es de 3 días.
  - La duración promedio en el Resort Hotel es de 4 días y medio.

```
#4) ¿Cuál es la duración promedio de las estancias por tipo de hotel?
data_hotel_limpio$stays_in_week_nights <- as.numeric(as.character(data_hotel_limpio$stays_in_week_nights))
data_hotel_limpio$stays_in_weekend_nights <- as.numeric(as.character(data_hotel_limpio$stays_in_weekend_n

data_hotel_limpio %>%
  mutate(estancia_total = stays_in_week_nights + stays_in_weekend_nights) %>%
  group_by(hotel) %>%
  summarise(duracion_promedio = mean(estancia_total, na.rm = TRUE))

data_hotel_limpio %>%
  mutate(estancia_total = stays_in_week_nights + stays_in_weekend_nights) %>%
  group_by(hotel) %>%
  summarise(duracion_promedio = mean(estancia_total, na.rm = TRUE)) %>%
  ggplot(aes(x = hotel, y = duracion_promedio, fill = hotel)) +
  geom_col(width = 0.6) +
  labs(title = "Duración Promedio de Estancia por Tipo de Hotel",
       x = "Tipo de Hotel",
       y = "Duración Promedio (noches)") +
  theme_minimal()
```

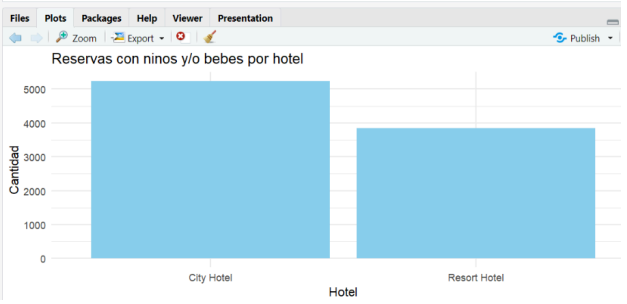
```
> #4) ¿Cuál es la duración promedio de las estancias por tipo de hotel?
> data_hotel_limpio$stays_in_week_nights <- as.numeric(as.character(data_hotel_limpio$stays_in_week_nights))
> data_hotel_limpio$stays_in_weekend_nights <- as.numeric(as.character(data_hotel_limpio$stays_in_weekend_night
s))
> data_hotel_limpio %>%
+   mutate(estancia_total = stays_in_week_nights + stays_in_weekend_nights) %>%
+   group_by(hotel) %>%
+   summarise(duracion_promedio = mean(estancia_total, na.rm = TRUE))
# A tibble: 2 × 2
  hotel        duracion_promedio
  <chr>                    <dbl>
1 City Hotel                3.15
2 Resort Hotel              4.42
```



- ¿Cuántas reservas incluyen niños y/o bebés?
  - En el City Hotel, las reservas que incluyen niños y/o bebés se sitúan aproximadamente entre 5,000 y 5,300. Por otro lado, en el Resort Hotel, estas reservas oscilan entre 3,700 y 4,000. Esto sugiere que las familias con menores tienden a preferir ligeramente el City Hotel.

```
#5) ¿Cuántas reservas incluyen niños y/o bebés?
reservas_con_ninos <- data_hotel_limpio %>%
  filter(as.numeric(children) > 0 | as.numeric(babies) > 0)

ggplot(reservas_con_ninos, aes(x = hotel)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Reservas con ninos y/o bebes por hotel", x = "Hotel", y = "Cantidad") +
  theme_minimal()

cat("Reservas con ninos o bebes:", nrow(reservas_con_ninos), "\n")


ggplot(reservas_con_niños_bebes, aes(x = hotel)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Reservas con ninos y/o bebes por hotel", x = "Hotel", y = "Cantidad") +
  theme_minimal()
```

```
> #5) ¿Cuántas reservas incluyen niños y/o bebés?
> reservas_con_ninos <- data_hotel_limpio %>%
+   filter(as.numeric(children) > 0 | as.numeric(babies) > 0)
> ggplot(reservas_con_ninos, aes(x = hotel)) +
+   geom_bar(fill = "skyblue") +
+   labs(title = "Reservas con ninos y/o bebes por hotel", x = "Hotel", y = "Cantidad") +
+   theme_minimal()
> cat("Reservas con ninos o bebes:", nrow(reservas_con_ninos), "\n")
Reservas con ninos o bebes: 9089
>
```

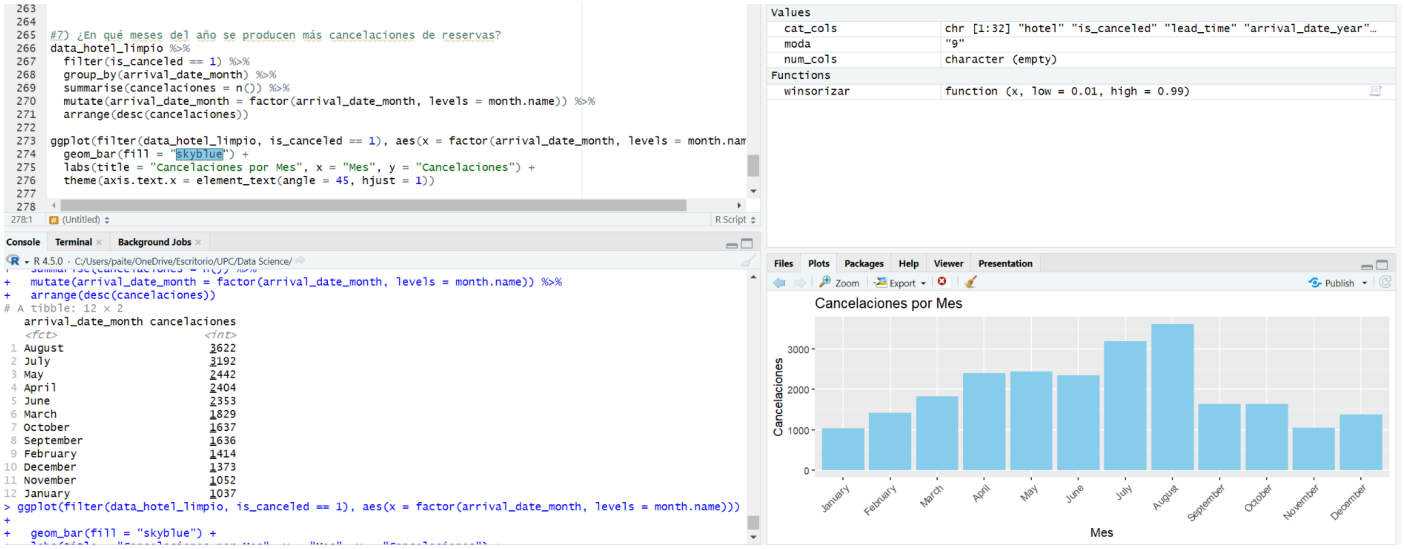


- ¿Es importante contar con espacios de estacionamiento?
  - El gráfico muestra como las únicas reservas canceladas se dieron en hoteles sin ningún estacionamiento, por ende es importante contar con al menos un estacionamiento para evitar la pérdida de clientes.

```
#6) ¿Es importante contar con espacios de estacionamiento?

ggplot(data_hotel_limpio, aes(x = as.factor(required_car_parking_spaces), fill = as.factor(is_canceled))) +
  geom_bar(position = "fill") +
  labs(title = "Relacion entre estacionamiento y cancelacion", x = "Espacios requeridos", y = "Proporcion") +
  scale_fill_manual(values = c("skyblue", "red"), name = "¿Cancelado?", labels = c("No", "Si")) +
  theme_minimal()
```

```
labs(title = "Reservas con ninos y/o bebes por hotel", x = "Hotel", y = "Cantidad") +
theme_minimal()
t("Reservas con ninos o bebes:", nrow(reservas_con_ninos), "\n")
rvas con ninos o bebes: 9089
plot(data_hotel_limpio, aes(x = as.factor(required_car_parking_spaces), fill = as.factor(is_canceled))) +
geom_bar(position = "fill") +
labs(title = "Relacion entre estacionamiento y cancelacion", x = "Espacios requeridos", y = "Proporcion") +
scale_fill_manual(values = c("skyblue", "red"), name = "¿Cancelado?", labels = c("No", "Si")) +
theme_minimal()
```

- ¿En qué meses del año se producen más cancelaciones de reservas?
  - Se puede notar, según los resultados en la consola, que agosto es el mes con mayor cantidad de cancelaciones a lo largo del año..



- Pregunta adicional: ¿Influye el canal de distribución en las cancelaciones?
  - Sí, el canal de distribución tiene un impacto importante en la tasa de cancelación.
  - Los canales Corporate, Direct y GDS son más confiables, mientras que los canales como TA/TO y especialmente los Undefined presentan mayores riesgos de cancelación.

Respuestas de las preguntas dentro del código:

#RESPUESTA A LAS PREGUNTAS DEL DOCUMENTO

data_hotel_limpio <- read.csv("hotel_bookings_limpio.csv", stringsAsFactors = FALSE)

Convertimos la columna de fecha
data_hotel_limpio$arrival_date <- ymd(data_hotel_limpio$arrival_date)

Creamos columnas adicionales
data_hotel_limpio$mes_llegada <- month(data_hotel_limpio$arrival_date, label = TRUE, abbr = TRUE)
data_hotel_limpio$duracion_estancia <- as.numeric(data_hotel_limpio$stays_in_weekend_nights) +
  as.numeric(data_hotel_limpio$stays_in_week_nights)

#1) ¿Cuantas reservas se realizaron por tipo de hotel?¿Que tipo de hotel prefiere la gente?
table(data_hotel_limpio$hotel)
prop.table(table(data_hotel_limpio$hotel)) * 100

```
ggplot(data_hotel_limpio, aes(x = hotel)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Reservas por tipo de hotel", x = "Tipo de hotel", y = "Cantidad de reservas") +
  theme_minimal()

#2) ¿Está aumentando la demanda con el tiempo?
data_hotel_limpio %>%
  group_by(arrival_date) %>%
  summarise(reservas = n()) %>%
  ggplot(aes(x = arrival_date, y = reservas)) +
  geom_line(color = "skyblue") +
  labs(title = "Evolución de la demanda con el tiempo", x = "Fecha", y = "Reservas") +
  theme_minimal()


#3) ¿Cuáles son las temporadas de reservas (alta, media, baja)?
data_hotel_limpio %>%
  group_by(arrival_date_month) %>%
  summarise(reservas = n()) %>%
  arrange(desc(reservas))

data_hotel_limpio$arrival_date_month <- factor(data_hotel_limpio$arrival_date_month,
                                               levels = month.name)

ggplot(data_hotel_limpio, aes(x = arrival_date_month)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Reservas por Mes", x = "Mes", y = "Reservas") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))


#4) ¿Cuál es la duración promedio de las estancias por tipo de hotel?
data_hotel_limpio$stays_in_week_nights <-
as.numeric(as.character(data_hotel_limpio$stays_in_week_nights))
data_hotel_limpio$stays_in_weekend_nights <-
as.numeric(as.character(data_hotel_limpio$stays_in_weekend_nights))

data_hotel_limpio %>%
  mutate(estancia_total = stays_in_week_nights + stays_in_weekend_nights) %>%
  group_by(hotel) %>%
  summarise(duracion_promedio = mean(estancia_total, na.rm = TRUE))

data_hotel_limpio %>%
  mutate(estancia_total = stays_in_week_nights + stays_in_weekend_nights) %>%
  group_by(hotel) %>%
  summarise(duracion_promedio = mean(estancia_total, na.rm = TRUE)) %>%
  ggplot(aes(x = hotel, y = duracion_promedio, fill = hotel)) +
  geom_col(width = 0.6) +
```

```
  labs(title = "Duración Promedio de Estancia por Tipo de Hotel",
       x = "Tipo de Hotel",
       y = "Duración Promedio (noches)") +
  theme_minimal()



#5) ¿Cuántas reservas incluyen niños y/o bebés?
reservas_con_ninos <- data_hotel_limpio %>%
  filter(as.numeric(children) > 0 | as.numeric(babies) > 0)

ggplot(reservas_con_ninos, aes(x = hotel)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Reservas con ninos y/o bebes por hotel", x = "Hotel", y = "Cantidad") +
  theme_minimal()

cat("Reservas con ninos o bebes:", nrow(reservas_con_ninos), "\n")



ggplot(reservas_con_niños_bebes, aes(x = hotel)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Reservas con ninos y/o bebes por hotel", x = "Hotel", y = "Cantidad") +
  theme_minimal()


#6) ¿Es importante contar con espacios de estacionamiento?

ggplot(data_hotel_limpio, aes(x = as.factor(required_car_parking_spaces), fill =
as.factor(is_canceled))) +
  geom_bar(position = "fill") +
  labs(title = "Relacion entre estacionamiento y cancelacion", x = "Espacios requeridos", y =
"Proporcion") +
  scale_fill_manual(values = c("skyblue", "red"), name = "¿Cancelado?", labels = c("No", "Si")) +
  theme_minimal()


#7) ¿En qué meses del año se producen más cancelaciones de reservas?
data_hotel_limpio %>%
  filter(is_canceled == 1) %>%
  group_by(arrival_date_month) %>%
  summarise(cancelaciones = n()) %>%
  mutate(arrival_date_month = factor(arrival_date_month, levels = month.name)) %>%
  arrange(desc(cancelaciones))

ggplot(filter(data_hotel_limpio, is_canceled == 1), aes(x = factor(arrival_date_month, levels =
month.name))) +
  geom_bar(fill = "skyblue") +
  labs(title = "Cancelaciones por Mes", x = "Mes", y = "Cancelaciones") +
```

```
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

#8) Pregunta adicional: ¿Influye el canal de distribución en las cancelaciones?
ggplot(data_hotel_limpio, aes(x = distribution_channel, fill = as.factor(is_canceled))) +
  geom_bar(position = "fill") +
  labs(title = "Cancelación por canal de distribución", x = "Canal", y = "Proporción") +
  scale_fill_manual(values = c("skyblue", "red"), name = "¿Cancelado?", labels = c("No", "Sí")) +
  theme_minimal()
```

4. CONCLUSIONES

a. Conclusiones basadas en el análisis:
   i. La mayoría de los clientes prefieren el City Hotel, con 53,422 reservas, frente a las 33,522 del Resort Hotel. Por lo que podemos decir que existe una preferencia por los hoteles en ubicaciones urbanas, probablemente por la cercanía a más servicios.
   ii. De los datos se observa una demanda con sostenido crecimiento en el tiempo, y con esto la evolución del negocio hotelero. Lo que muestra que las estrategias de negocios, ajustes en logística o mejora de la atención, aplicadas por los hoteles están funcionando.
   iii. Durante los meses de verano(Junio, Julio,Agosto) hay una alta demanda, mientras que invierno y primavera/otoño por otro lado muestran una demanda baja y media respectivamente. Las promociones y estrategias se ajustan a la estación del año.
   iv. La duración de las estancias varían dependiendo el tipo de hotel, en los resorts son más prolongadas, 4.5 días en promedio, y en city hotel 3 días en promedio.
   v. El city hotel atrae mayor cantidad de familias, si bien los resort también las reciben, probablemente en el city hotel las familias encuentran más accesibilidad para viajes cortos, esto se ve en que el city hotel has recibido entre 5000 y 5300 reservas versus resort que recibe entre 3700 y 400 reservas de familias.
   vi. La alta demanda que se ve en agosto también puede generar cancelaciones en las reservas debido a cambios de planes o sobre reservas.
   vii. Los canales de distribución influyen en las cancelaciones, se ha notado que canales como Corporate, Direct Y GDS tienen bajas tasas de cancelación mientras que TA/TO y los undefined presentan mayor riesgo.
b. Recomendaciones
   i. Claridad en las Visualizaciones
      1. Todos los gráficos deben llevar títulos claros, etiquetas en ejes y leyendas explicativas.

2. Si se comparan variables por tipo de hotel (City vs Resort), usar colores consistentes y separar los gráficos si es necesario para facilitar la interpretación.
3. Visualizaciones sugeridas:

    a. Barras comparativas por tipo de hotel y por canal de distribución.

    b. Series temporales para observar la evolución mensual de reservas y cancelaciones.

    c. Mapas de calor o gráficos de líneas para identificar temporadas altas/bajas.

5. **ANEXO:**
   Github**:** https://github.com/Maserattew/1ACC0216--TB1-2025-1/tree/main