Reporte de Entendimiento Inicial de Datos

1. Descripción General del Dataset

El *dataset* utilizado contine la información de reservas realizadas en hoteles entre 2015 y 2017, cuenta con 33 atributos y un total de 58 895 registros, tras realizar la limpieza de los datos se redujo el número a 56 501 registros válidos (cuentan con la información mínima completa).

2. Tipos de Datos Contenidos

El dataset cuenta con 19 variables de tipo numérico entero, 12 categóricas y 1 variable numérica decimal. De acuerdo con su tipo están distribuidas cómo:

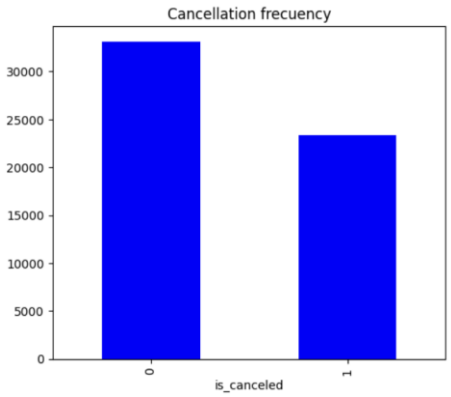
- **Cuantitativa discreta (12):** *lead_time, stays_in_weekend_nights, stays_in_week_nights, adults, children, babies, previous_cancellations, previous_bookings_not_canceled, booking_changes, days_in_waiting_list, required_car_parking_spaces, total_of_special_requests*
- **Cuantitativa continua (1):** *adr*
- **Cualitativa nominal (15):** *hotel, meal, country, market_segment, distribution_channel, reserved_room_type, assigned_room_type, deposit_type, customer_type, reservation_status, agent, company, is_canceled, is_repeated_guest, arrival_date_month*
- **Cualitativa ordinal (4):** *arrival_date_year, arrival_date_week_number, arrival_date_day_of_month, reservation_status_date*

3. Top 5 Atributos Más Importantes

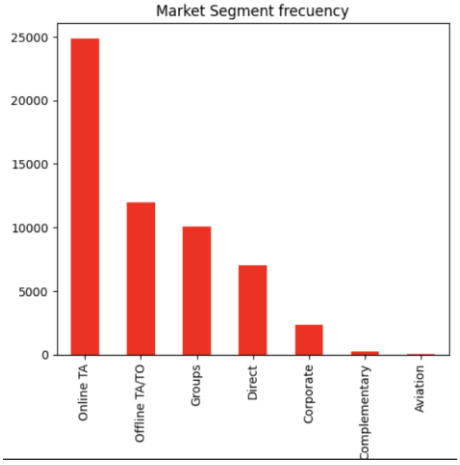
Para el entendimiento inicial, se seleccionaron cinco atributos considerados relevantes por su papel en el comportamiento de la demanda

3.1 Atributos categóricos

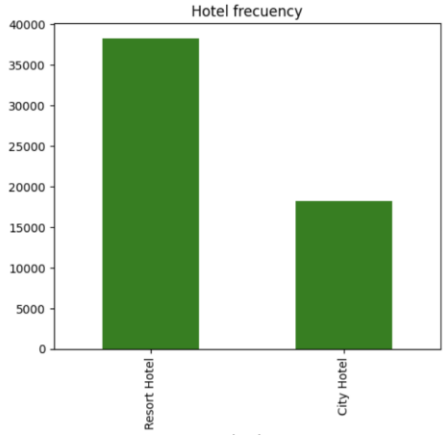
1. **is_canceled:** Indica si la reservado fue o no cancelada (Binario). El análisis univariado muestra una frecuencia de cancelación normalizada del 41.3%. Dada su prevalencia y el impacto de esta métrica en la ocupación será utilizada como variable principal/objetivo.



2. **market_segment:** representa el canal desde el cuál se genera la reserva, esta variable permitirá entender de qué manera los distintos canales influyen el comportamiento de los clientes. Existen 7 valores posibles para esta variable como líder absoluto Online TA representando el 43.9% de las reservas, doblando al siguiente grupo Offline TA/TO (21.2%).

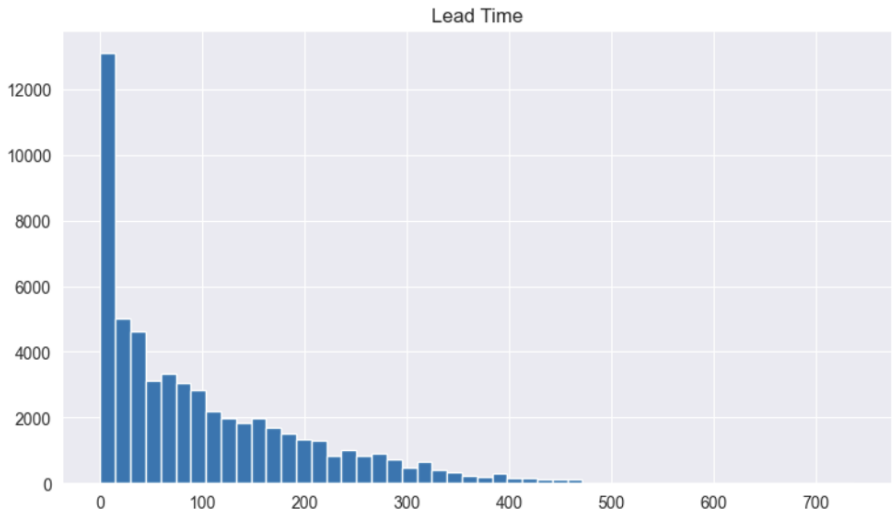


3. **hotel:** representa el tipo de hotel sobre el cuál se realizó la reserva, esta variable permitirá entender cómo las dinámicas asociadas a cada tipo de hotel pueden afectar la demanda. Existen dos valores posibles Resort Hotel con dos tercios de las reservas y City Hotel con el excedente.



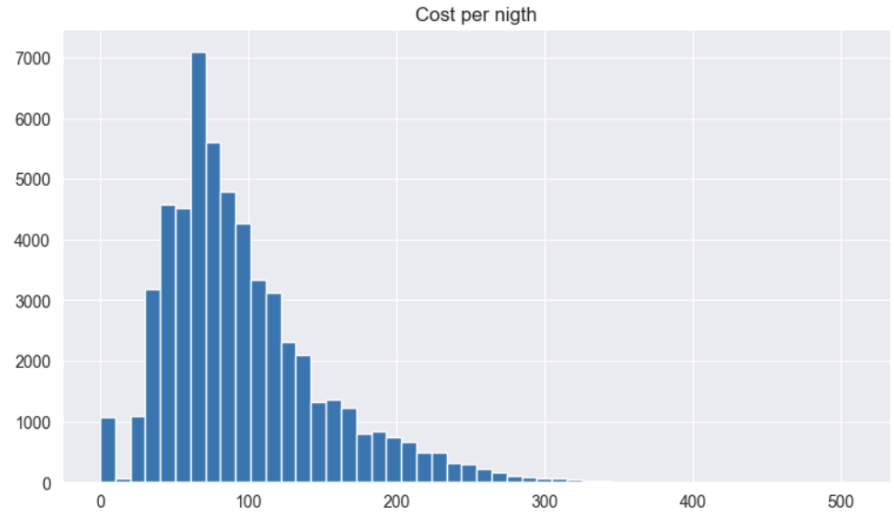
3.1 Atributos cualitativos

4. **lead_time:** mide el tiempo (en días) de anticipación entre la reserva y la fecha de llegada, se utilizará para entender si existe alguna relación entre esta variable y la tasa de cancelación. Presenta un valor promedio de 100, media de 69 y desviación estándar de 101; adicionalmente, a partir de los cuartiles se pueden concluir que la mayoría de las reservas se hacen a corto plazo (25% de las reservas con al menos 2 semanas y 50% con 7 semanas), sin embargo, en el 25% de las reservas (Q4) se hacen con más de 5 meses de anticipación. Estos valores reflejan que en las reservas existen comportamientos tanto de planificaciones a largo plazo como decisiones de última hora.



Por demás, la métrica a cuenta con asimetría (skewness) de 1.2 que nos permite concluir que la mayoría de las reservas que se realizan con "poca" (relativa) anticipación. La distribución de este atributo cuenta con kurtosis cercana a 1 lo que indica una mayor presencia de valores extremos/atípicos.

5. **adr:** mide el coto por noche promedio asociado a la reserva. Presenta un valor promedio de 96 con media de 84 y una desviación estándar de 54. Aunque la mayoría de los valores se concentran en



Por último, la métrica cuenta con asimetría (skewness) de 1.23 dado que la mayoría de las reservas presenta valores bajos de adr. Su distribución presenta una Kurtosis de 1.99 que indica una alta presencia de valores atípicos.