Trabajo Práctico 1
Reservas de Hotel

[75.06/95.58] Organización de Datos
Primer cuatrimestre de 2023

ALUMNO	PADRON	CORREO
BARTOCCI, Camila	105781	cbartocci@fi.uba.ar
HERBAS, Enzo	103747	jeherbas@fi.uba.ar
PATIÑO, Franco	105126	fpatino@fi.uba.ar

Sobre el análisis de los tipos de datos y sus frecuencias

En primer lugar, aquellas columnas que son fechas, las pasamos a tipo datetime ya que nos interesa poder ordenarlas y trabajar matemáticamente con ellas. Pasamos las columnas que especifican datos booleanos de int a object para poder trabajarlas como categóricas. Por otro lado, decidimos trabajar la variable 'lead_time' en meses en vez de en días para facilitar la interpretación de los datos al trabajar con una unidad que se ajuste mejor a sus valores.
También observamos que 'country' tenía muchas categorías con pocos registros, para tratarlas definimos un umbral de frecuencia baja del 5 % sobre el total de los datos, estas categorías las englobamos en una nueva llamada 'others', reduciendo así la cantidad de categorías totales.
Modificamos la columna 'days_in_waiting_list', ya que los valores se concentraban mayoritariamente en el valor 0. Como la consideramos importante para el análisis de cancelación de reservas, la conservamos transformandola en una booleana más fácil y relevante de analizar a nuestro parecer, 'more_than_zero_days_in_waiting_list'.
En algunos casos agrupamos categorías que tenían frecuencias muy bajas, como en 'deposit_type' y en 'customer_type'.

Sobre las columnas unificadas, las generadas y las eliminadas

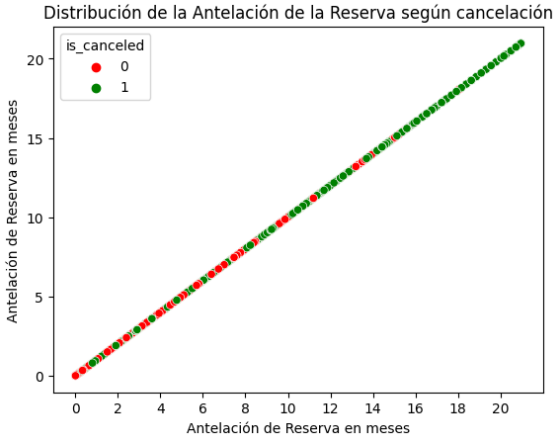
Eliminamos la columna 'id' al considerar que una identificación no aporta nada para el análisis del problema.
Por otro lado, consideramos que tener los huéspedes clasificados en 'adults', 'children' y 'babies' no era tan relevante como tener el total de huéspedes en una única variable 'total_guests'. Tras un análisis de las medidas de resúmen, de las distribuciones de estas cuatro variables y de sus relaciones con 'is_canceled', concretamos la unificación de esas tres columnas en 'total_guests'.
En cuanto a la columna 'required_car_parking_spaces' no nos parecía relevante para el análisis y, luego de analizar su distribución y ver que la gran mayoría de sus datos se concentraba en una única categoría, decidimos eliminarla. Para 'meal', consideramos que dadas las categorías especificadas, los hoteles ofrecen todas las opciones de comida por momento del día, por lo que no nos parecía relevante tenerla en cuenta. Por estos motivos y tras analizar su distribución y ver que la gran mayoría de datos caían en una misma categoría, decidimos eliminarla.
Decidimos también unificar las columnas referentes a día, mes y año en una única 'arrival_date', eliminando así estas tres columnas y aquella que indicaba el número de semana.
También nos pareció interesante generar una columna 'season' para analizar las reservas agrupadas por estación del año, y otra columna 'room_type_match' que compara si 'assigned_room_type' y 'reserved_room_type' son iguales.

Sobre los datos faltantes

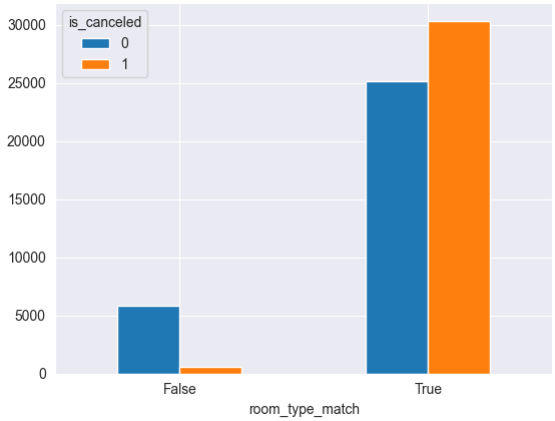
A partir de un análisis gráfico, detectamos que cuatro columnas tenían algunos datos nulos.
La variable 'company' era la que más datos nulos tenía en el dataset, alrededor del 95 % del total. En este caso, decidimos eliminar la columna ya que creemos que es un porcentaje muy alto de datos nulos como para reconstruirlos mediante algún método de imputación. En cuanto a 'agent', teníamos un porcentaje cercano al 13 % de datos nulos, además de una importante cantidad de categorías. Dado que consideramos que es una variable interesante para la predicción del target y que no es un porcentaje alto de nulos, optamos por hacer un análisis más profundo. Notamos que esta variable podría tener una relación con las variables 'market_segment' y 'distribution_channel', bajo el supuesto de que la nulidad de 'agent' podría estar relacionada al segmento de mercado y/o al canal de distribución de la reserva, mediante la categoría 'TA/TO'. Decidimos estudiar los heatmaps de ambas variables contra 'agent', y observamos que en ambos casos los valores más cálidos se daban cuando el agente no era nulo y el tipo era 'TA/TO'. A partir de eso, concluimos que la nulidad del agente está relacionada al segmento de mercado y al canal de distribución, por lo que vimos más beneficioso transformar 'agent' a una variable booleana 'agent_specified'. Con esto logramos además facilitar el análisis de la misma al reducir la cantidad de categorías a sólo dos.

Con respecto a 'country', observamos que había un porcentaje del 0.35 % de nulos. Al ser un porcentaje bajo y ver que la distribución no sufría cambios significativos, decidimos imputarlos englobandolos dentro de la categoría 'others'.
El caso de 'children' tenía apenas 0.0064 % de datos nulos. Sencillamente optamos por imputarlos reemplazandolos por el valor 0. Nuevamente, la distribución no sufrió cambios importantes.

Dos visualizaciones relevantes de nuestro análisis



Consideramos que este gráfico muestra una correlación lineal importante entre la antelación de la reserva en meses y el estado de cancelación. Se observa claramente que, cuando se reserva con más de 10 meses de antelación, hay una tendencia a cancelar la reserva.



En este gráfico vemos que, si bien se asignó el cuarto reservado por el cliente, la frecuencia de cancelación es mucho mayor que cuando se asigna un cuarto distinto al reservado.