Materia: 75.06/95.58 - Organización de Datos
Cuatrimestre: 1° cuatrimestre 2023
Grupo: 27

Padrón	Apellido y Nombre
108405	Porro, Joaquín
97538	Bordón Villavicencio, Fernando Nahuel

Algunas Variables Irrelevantes

arrival_date_week_number Puede ser deducidas con *arrival_date_year* *arrival_date_month* y *arrival_date_day_of_month*

---

stays_in_weekend_nights y *stays_in_week_nights*
Podrían ser deducidas con una nueva variable `*stays*` (que indicaría cuantos dias se quedaría) junto con *arrival_date_year*, *arrival_date_month* y *arrival_date_day_of_month*

Primero, la columna "*id*" parece ser una simple identificación numérica de cada registro y, por lo tanto, es poco probable que proporcione información útil para el análisis.

Además, las columnas "*reserved_room_type*" y "*assigned_room_type*" pueden proporcionar información similar, ya que ambas indican el tipo de habitación reservada o asignada al huésped. Si el análisis solo necesita una de estas variables, la otra podría ser redundante.

Análisis Exploratorio

Las cosas más importantes que vimos en el análisis son:

- Sólo hay valores nulos en children (`4`), country (`221`), agent (`7890`) y company (`58761`) en el dataset `hotels_train`, y en country (`95`), agent (`3363`) y company (`25218`) en el dataset `hotels_test`.

En `hotels_train` En decidimos hacer dropna() a todas las filas que tengan children nulo, dado que son pocas filas y no se pierden muchos datos. Para las de country: Para las country nulas, hicimos un fillna() con valor "Unknown" para indicar que no se conoce el country, y así evitamos eliminar demasiados datos. Para agent y company hicimos algo similar con fillna('None'), indicando que no tiene agente o compañía. En `hotels_test` hicimos lo mismo, solo que no hicimos nada con los children porque acá no hay valores children nulos.
Luego, eliminamos filas duplicadas.

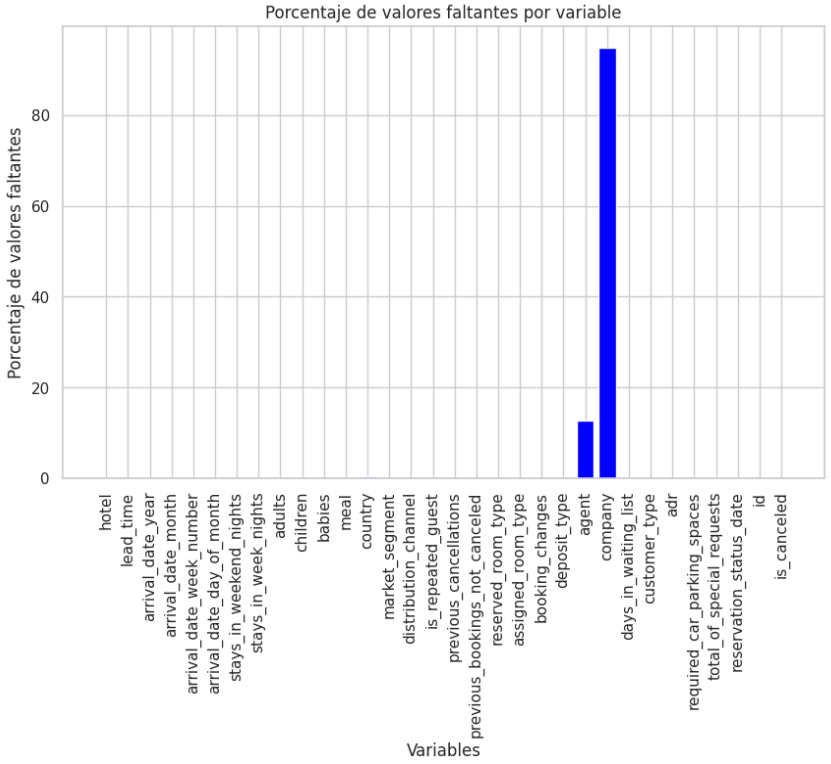
Análisis de Valores Atípicos

Después hicimos algunos boxplots para ver gráficamente algunos valores atípicos, y luego los buscamos de forma univariada y de forma multivariada
A la derecha está la cantidad de valores atípicos que encontramos de forma univariada en cada columna, y con el análisis multivariado encontramos 11160 valores entre todas las columnas

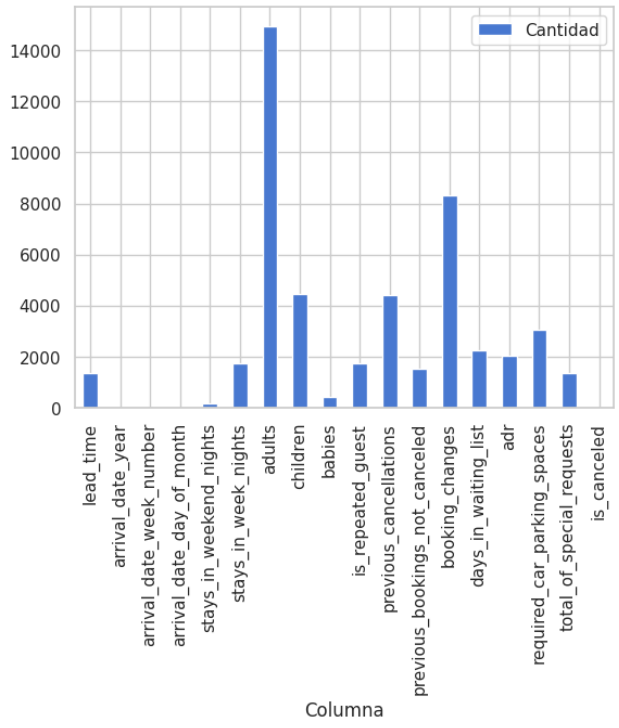
	Valores_atipicos
lead_time	1369
arrival_date_year	0
arrival_date_week_number	0
arrival_date_day_of_month	0
stays_in_weekend_nights	144
stays_in_week_nights	1724
adults	14950
children	4452
babies	428
is_repeated_guest	1733
previous_cancellations	4394
previous_bookings_not_canceled	1538
booking_changes	8317
days_in_waiting_list	2235
adr	2025
required_car_parking_spaces	3072
total_of_special_requests	1358
is_canceled	0

Con respecto a estos valores outliers, pensamos que lo mejor es reemplazar a los que son un error (como por ejemplo hay un valor de address que es 96.67, lo cual no es posible), y dejar los que no lo son cómo están.

Gráficos



En este gráfico se puede ver que faltan más del 80% de los datos de company y poco menos del 20% de los datos de agent, lo que justifica que no eliminemos todas esas filas, dado que perderíamos gran parte del dataset



Este otro gráfico muestra los valores atípicos encontrados de forma univariada