Checkpoint 1 - Grupo 03

Análisis Exploratorio

En el trabajo utilizamos un dataset de reservas de hotel. El mismo tiene un total de *61913* filas (que corresponden a la cantidad de reservas) y *31* columnas. Cabe destacar que cada fila tiene un id y confirmamos que el mismo no se repite. Entre esas columnas algunas de las más destacadas son:

- **is_canceled:** es una variable que asumimos categórica, y representa si la reserva fue (*1*) o no (*0*) cancelada; para este tp será nuestro *target* a predecir.
- **adults, children y babies** son variables cuantitativas discretas que usaremos para determinar para cuantas personas se hizo la reserva. A partir de estas columnas después creamos una nueva, **people**.

Contando los valores que toma **people**, vemos que los valores mayores a 5 son el *0.00330%* del total. Por tanto resolvimos establecer que cada reserva tiene máximo 5 personas con al menos 1 adulto.

Preprocesamiento de Datos

Columnas eliminadas:

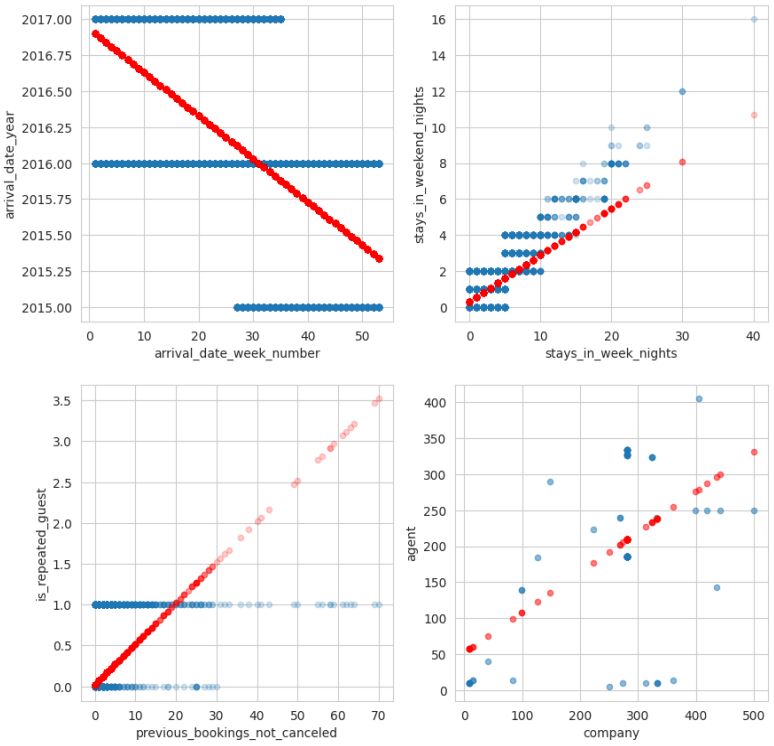
Se eliminó la columna **company**, ya que descubrimos que el *94.909%* de los valores de esta columna eran nulos (***nan***).

Correlaciones detectadas:

Notamos correlaciones sobresalientes entre los siguientes pares de columnas, junto a sus correlaciones de Pearson:

1. **arrival_date_year** y **arrival_date_week_number** (*-0.540542*)
2. **stays_in_weekend_nights** y **stays_in_week_nights** (*0.48871*)
3. **previous_bookings_not_canceled** e **is_repeated_guest** (*0.40603*)
4. **agent** y **company** (*0.514969*)

Gráficos de Dispersión:
Variables con correlación Sobresalientes



Columnas recodificadas:

Variables como **children** y **agent**, que eran de tipo *float*, fueron casteados a tipo *int* ya que todos sus datos terminaban en '*.0*' , y no tenía sentido por ejemplo tener "2 niños y medio".

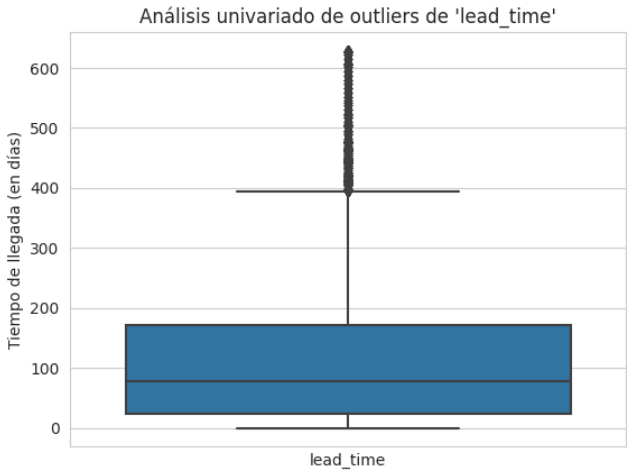
Valores atípicos:

Se encontraron *outliers* en casi todas las columnas, pero resaltan los casos de algunas de las siguientes: **lead_time**,**people**, **adr**, **previous_bookings_not_canceled**, o **total of special requests**.
La gran mayoría presentaba *outliers* pero no necesariamente valores atípicos: esto se debe a lo compactos que estaban la mayoría de los datos.

Análisis Univariados:

lead_time: En esta variable se encuentran muchos outliers superiores, con muchísimas reservas que esperan más de 400 días.

adr: En esta variable utilizamos el método *z-score* modificado para detectar los *outliers* superiores. Para este método usamos una regla de oro: valores mayores a *3.5* se consideran *outliers*.
total_of_special_requests: En esta variable los valores se concentran entre *0* y *1*, los *outliers* como mucho llegan a *5*. Como estos valores no son muy descabellados, decidimos dejar esta columna como estaba.



Análisis Multivariados:

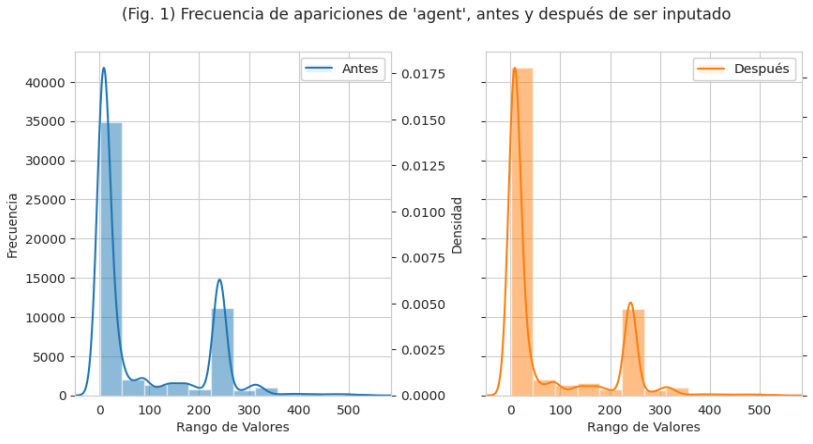
stays_in_weekend_nights y **stays_in_week_nights**: En su respectivo análisis de correlación, notamos que habían entradas con *0* en ambas variables: éstos son valores atípicos multivariados. Su porcentaje resulta ser menos del *0,5%* del total así que resolvemos eliminarlos.
is_repeated_guest y **previous_booking_not_canceled**: Decidimos borrar los casos donde figura que es la primera vez que un cliente viene pero tiene una o más reservas previas no canceladas, lo cual no tiene sentido.

También notamos que las filas con *0* adultos son menos del 1%, así que las eliminamos. Algo similar pasa con las filas con más de 5 en la variable **people**, también eliminadas por ser demasiado pocas y no tener sentido.
Además borramos los valores de **adr** que sean *0* o negativos ya que carece de lógica no pagar la reserva o que el hotel te pague.

Valores faltantes:

Nombre de la Columna	Porcentaje de Valores Faltantes	Resolución
company	*94.909%*	Eliminamos la columna entera.
agent	*12.744%*	Imputamos con valor más frecuente (9)
country	*0.357%*	Borramos sólo las filas afectadas.
children	*0.006%*	Borramos sólo las filas afectadas.

Visualizaciones



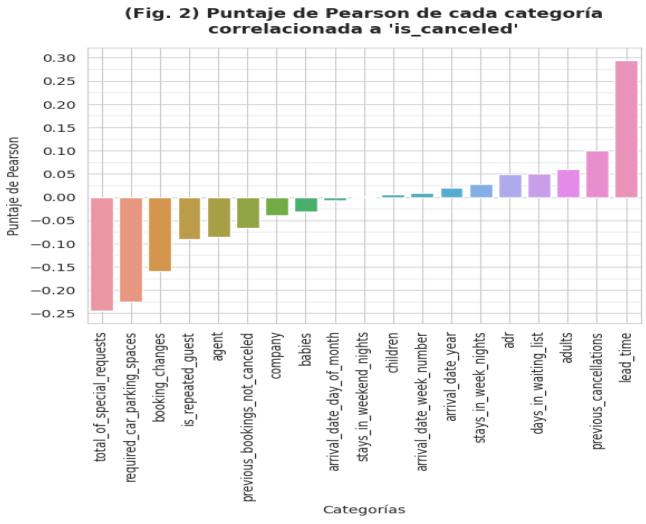


Fig. 1: Breve comparación de las frecuencias de apariciones de ***agent*** después de ser imputado.
Fig. 2: Correlación de todas las columnas numéricas con el *target*. Notar que ninguna es particularmente mayor a $|0.3|$.

Tareas Realizadas

Integrante	Tarea
Franco Lighterman Reismann	Mantención del repositorio Armado de gráficos Imputación de Datos
Marcos García Neira	Análisis de Valores Faltantes Armado de Reporte
Martín Andrés Maddalena	Detección de Outliers Análisis de Valores Atípicos