Checkpoint 1 - Grupo 25

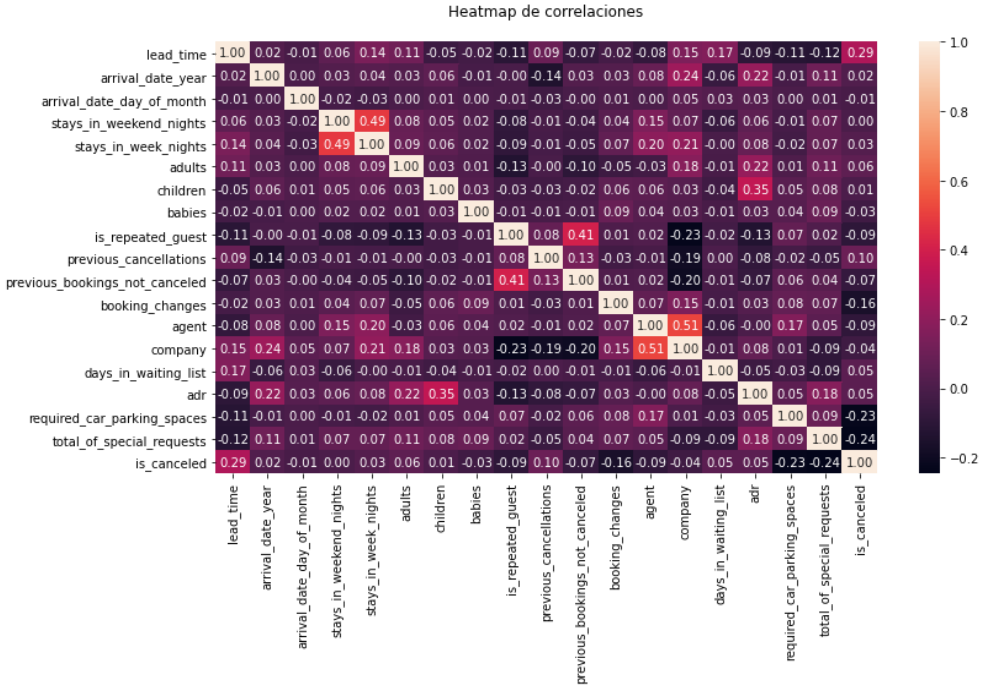
Análisis Exploratorio

El dataset tiene 31 columnas y 61913 filas.
A Priori, los features más destacables, son, teniendo en consideración la variable a predecir: **'previous_cancellations'** (cantidad de reservas anteriores que fueron canceladas), **'previous_bookings_not_canceled'** (cantidad de reservas anteriores que no fueron canceladas), **'is_repeated_guest'** (si el huésped es o no repetido), **'lead_time'** (tiempo de estadía), **'adr'** (tarifa promedio).

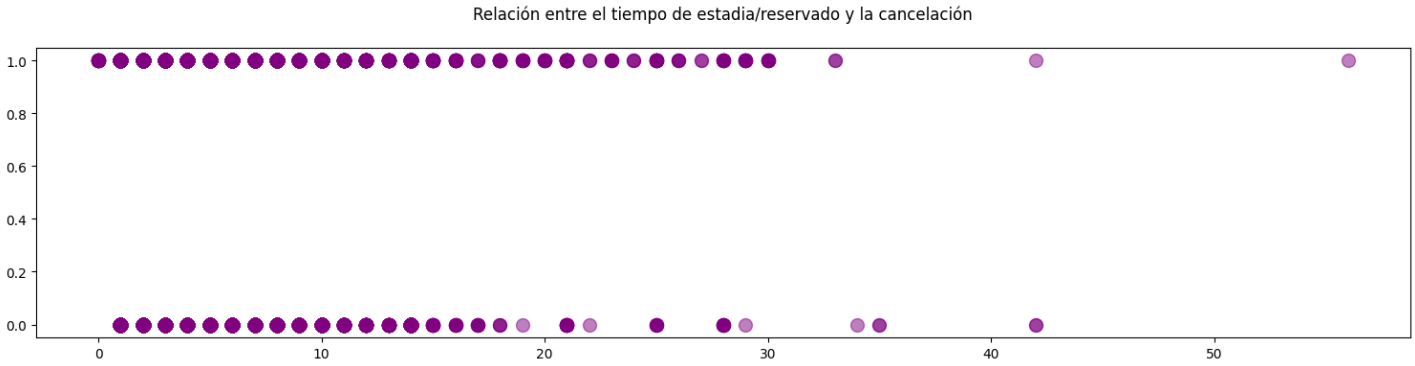
Preprocesamiento de Datos

1. Columnas eliminadas: **'id'**, como representa un código único que identifica cada reservación, no produce información sobre la cancelación de la misma. **'arrival_date_week_number'**, al eliminar esta columna no perdemos información ya que tenemos otras variables similares que representan las fechas de la reserva.
2. Correlaciones detectadas:
- **'children' - 'adr' (moderada)**. Coeficiente de Pearson: 0.350482.
- **'stays_in_weeknd_night' - 'stays_in_week_nights' (moderada)**. Coeficiente de Pearson: 0.488710.
- **'is_canceled' - 'lead_time' (moderada)**. Coeficiente de Pearson: 0.29.
- **'agent' - 'company' (fuerte)**. Coeficiente de Pearson: 0.51.
3. Columnas recodificadas: **'customer_type'**, **'deposit_type'**, **'meal'**, **'reserved_room_type'**, **'assigned_room_type', 'country', 'market_segment', 'distribution_channel'.** Esto es porque los algoritmos de aprendizaje automático requieren que las variables de entrada sean numéricas en lugar de categóricas
4. Valores atípicos:
- Univariadas: 'lead_time' y 'adr' (Z-Score).
- Multivariadas:
    - ☐ 'children', 'adults' y 'babies'.
    - ☐ 'stays_in_week_nights', 'stays_in_weekend_nights' y 'is_canceled'.
- Datos incongruentes: reservas con más de 8 personas, cancelaciones mayores a 5, requerimientos de espacios para los autos mayores a 2.
5. Valores faltantes: **'country'**, faltan 0.356952% de los datos (eliminados). **'children'**, faltan 0.006461% de los datos (eliminados). **'agent'** no se decide imputar ya que se reservó el hotel sin un agente de por medio (se rellena con -1). **'company'**, no se decide imputar ya que corresponde a pagos de reserva que no se realizaron a través de esta (se rellena con -1). El análisis entre estas columnas dice que un agente puede estar vinculado a más de una compañía. Por otra parte, fueron eliminadas las observaciones de compañías con alta tasa de reserva la cual era mayor al 90% de cancelaciones.

Visualizaciones



Elegimos el heatmap ya que nos muestra la correlación entre todas las variables incluyendo el target.



Elegimos este gráfico ya que nos parece una buena representación de una variable con respecto al target. Se puede observar que las reservas que están entre los días 20 y 30 se tienden a cancelarse mucho más.

Tareas Realizadas

Integrante	Tarea
Mariana Juarez Goldemberg	Exploración Inicial Visualización de los datos Datos Faltantes Valores atípicos univariados Armado de Reporte
Lisandro Roman	Datos Faltantes Armado de reporte Valores atípicos multivariados Creación de nuevas columnas
Miranda Marenzi	Visualización de los datos Datos Faltantes Análisis de valores incongruentes Recodificación de columnas Armado de Reporte