Checkpoint 1 - Grupo 24 Merequetengue 👍

Análisis Exploratorio

El dataset tiene aproximadamente 27000 registros y 31 columnas las cuales pudimos separar en variables cuantitativas, cualitativas y ordinales con el fin de realizar distintos métodos de procesamiento de información.
Luego de analizar los datos creemos que los features más destacables a la hora de predecir si va a ser cancelada la reserva son:"days in waiting list", creemos que una persona que esperó muchos días para que se le tome la reserva es más probable que cancele la misma.
"Previous cancellations", este dato nos dará una gran ayuda a la hora de saber si el inquilino es más propenso a cancelar. También "is repeated guest" creemos que un inquilino que ya estuvo en el hotel, es menos probable que cancele la reserva. Por otro lado también creemos que la cantidad de personas que adultos, niños y bebés que se quedan en la habitación es relevante ya que al ser más personas existe más probabilidad de que cancelen la reserva.

Preprocesamiento de Datos

1. Columnas eliminadas:

   Tomamos la decisión de no eliminar ninguna columna, ya que creemos que pueden ser útiles para futuros análisis.

2. Correlaciones detectadas:

   Luego de analizar el gráfico de correlaciones, no encontramos ningunas
   Variables correlacionadas fuertemente entre sí, ya sea positiva o negativamente.
   Si notamos que hay una leve correlación entre las variables 'previous bookings not canceled' y 'is repeated guest' con un 0.41 de correlación.

3. Columnas recodificadas:

   Hemos modificado 3 columnas, las primeras dos en ser modificadas fueron la columna "agent" y "company" debido a la gran cantidad de valores NaN encontrados, optamos por reemplazar todos estos valores NaN por el nombre "not applicable".
   Finalmente en la columna country también modificamos todos los valores NaN por "Undefined", para que sea más claro que no se sabe el país de procedencia.

4. Valores atípicos encontrados:

   Todos los que encontramos son univariados.

- Cantidad de adultos por habitación:
  Vemos extraño el hecho de que haya 20 personas o más en una habitación por lo que son valores atípicos como se muestran en el histograma. Solución planteada: Asumimos que los valores por encima de 8 fueron valores mal ingresados por lo que serán reemplazados con la mediana.

- Cantidad de bebés en cada familia:
  Al observar la cantidad de bebés por familia aparece un valor bastante atípico el cual nos indica que una familia tiene 10 bebés, entendemos que hubo un error al ingresar este valor.
  Solución: Cambiar el valor mal ingresado de la cantidad de bebés por la mediana.

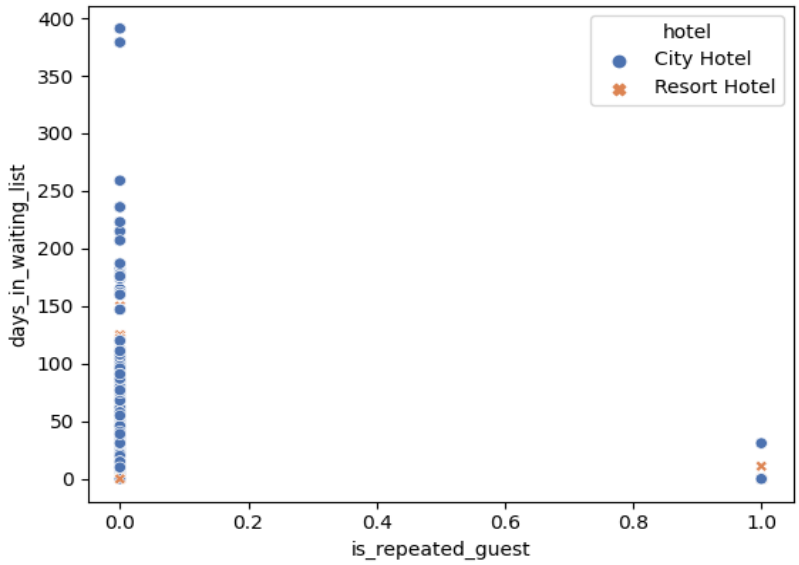
- ADR:
  Vemos que tenemos un valor exageradamente por encima de los demás, y es extremadamente atípico, por lo que también procedimos a cambiar este por el valor de la mediana.

- Otros valores atípicos que no han sido modificados: “Días en lista de espera”,”Cancelaciones previas”,”Fines de semana en el hotel”, Los primeros dos no han sido modificados por que consideramos que nos serán de gran utilidad a la hora de calcular el target. Y el tercero simplemente no nos pareció relevante como para modificarlo, el inquilino puede estar todos los fines de semana que desee siempre y cuando pague.

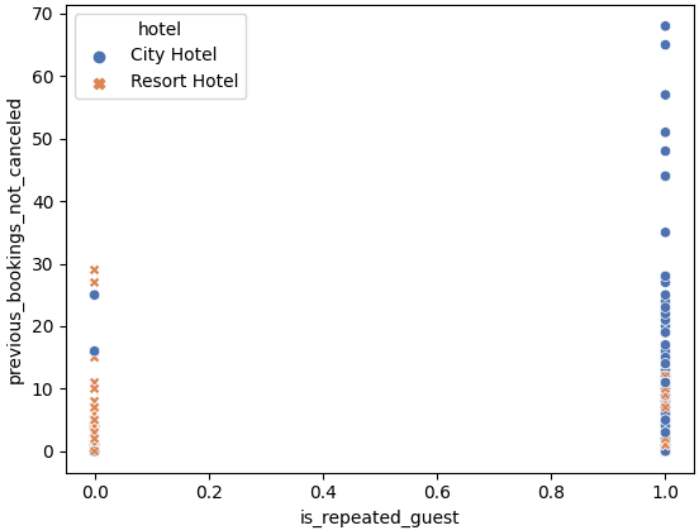
5. Valores faltantes:

   Las columnas con datos faltantes eran ‘country’, ‘agent’ y ‘company’, donde reemplazamos los valores faltantes con “not applicable” en las columnas ‘agent’ y ‘company’, y en la columna ‘country’ colocamos “undefined” para los registros sin países. La columna ‘country’

Visualizaciones



En el gráfico se puede visualizar como dependiendo del hotel y si es un cliente repetido reciben mayor o menor prioridad a la hora de ser atendidos.
Como podemos observar no existe una correlación entre ambas, teniendo aproximadamente 0 en el coeficiente de Pearson. Pero creemos que esta relación entre ambas seran util para calcular futuramente el target.



En el gráfico de la izquierda podemos ver como están correlacionadas las variables de si es un cliente repetido con cancelaciones previas.

Si se calcula el coeficiente de Pearson vemos que nos da un valor aproximado al 0.41.

Como la variable del eje X únicamente tiene valores binarios no podremos observar la existencia de una recta determinado la correlación

Sin embargo, creemos que va a ser útil para el cálculo del target.

Tareas Realizadas

Integrante	Tarea
Torraca Lautaro - 108813	Detección de Outliers Análisis de Correlaciones Armado de Reporte
Marco Tosi - 107237	Análisis de Correlaciones Análisis de Valores Faltantes Armado de Reporte
Gianluca Negrotti - 108184	Análisis de Valores Faltantes Exploracion inicial Armado de Reporte
Trabajamos todos en conjunto y debatiendo al mismo tiempo que avanzamos con el TP	