

Checkpoint 1 - Grupo 39

Análisis Exploratorio

Información del dataset:

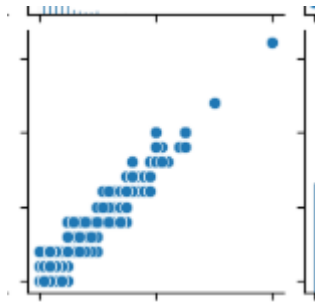
- Cantidad de registros: 61913
- Cantidad de columnas: 31
- Cantidad de variables numericas: 19
- Cantidad de variables Object (string): 11

Features destacadas:

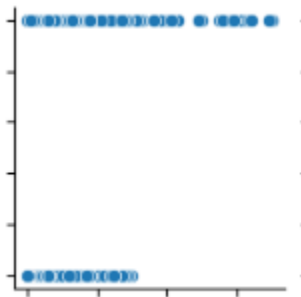
- Adults: representa la cantidad de adultos de la reserva, es de tipo numerico y fue muy importante para varios de los analisis de valores atipicos, ya que se relaciona de una manera logica con las demas variables
- Babies y Children: representan los bebes y chicos de la reserva, son de tipo numerico y en ambos casos pasa algo parecido que con adults se relaciona de manera logica con las demas variables y fue importante para la detección de outliers multivariados.
- Lead_time: representa el tiempo entre la reserva y el chek-in, es de tipo numerica y nos parece importante ya que el tiempo de espera de la llegada tiene mucho que ver con la cancelacion de la reserva, ya que en el medio pueden pasar varias cosas. Ademas la usamos para encontrar valores atipicos multivariados
- Is_repeated_guest: representa si la reserva es de un persona que ya ha hecho una reserva, es de tipo categorica y vimos que hay relacion entre esta y algunas de las variables (por ejemplo: previous_booking_not_canceled)

Preprocesamiento de Datos

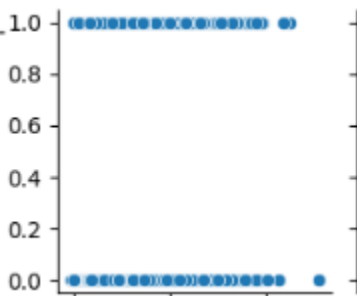
1. Columnas eliminadas: arrival_date_year, arrival_date_month, arrival_date_week_number, arrival_date_day_of_month, meal
Eliminamos esas columnas ya que las consideramos irrelevantes para nuestro análisis por tener una baja relación con el target
2. Correlaciones detectadas:
 - Stay_in_weekend_nights(eje y) y stays_in_week_nights(eje_x):
Valor de correlacion de Pearson: 0.45



- previous_booking_not_canceled (eje x) y is_repeated_guest(eje y):
Valor de correlacion de Pearson: 0.59

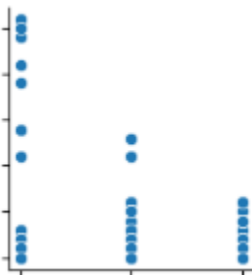


- is_canceled(eje y) y lead_time(eje x):
Valor de correlacion de Pearson: 0.28



Tienen una correlacion mediamente baja, pero nos interesa porque lead_time es la variable que tiene mas relacion con is_canceled, que es nuestra variable a predecir.

- previous_cancellations(eje y) y arrival_date_year(eje x):
Valor de correlacion de Pearson: -0.32



3. Columnas recodificadas: Ninguna

4. Valores atípicos:

Para encarar el análisis decidimos realizar el boxplot de cada variable cuantitativa para visualizar los outliers univariados (véase en la notebook) y luego identificarlos utilizando z score modificado (o z score en los casos en los que el MAD daba 0 y no nos permitía utilizar el z score modificado).

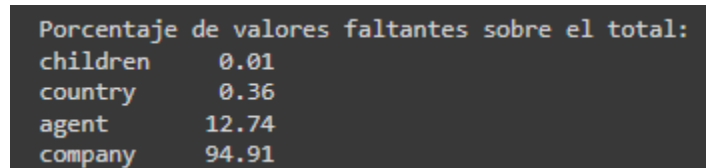
Adicionalmente, utilizamos nuestro criterio y conocimiento del área para, en algunos casos, mantener outliers que consideramos valores válidos y posibles (ej.: stays_in_week_nights, adults, children, entre otras) ya que al eliminarlos perderíamos información valiosa para predecir.

Con respecto a los multivariados, utilizamos un pairplot para visualizar los outliers y aplicamos mahalanobis para los casos en los que veíamos mayor cantidad de outliers.

Los gráficos asociados a dicho análisis pueden visualizarse en la notebook.

5. Valores faltantes:

Encontramos los siguientes datos faltantes:



Porcentaje de valores faltantes sobre el total:	
children	0.01
country	0.36
agent	12.74
company	94.91

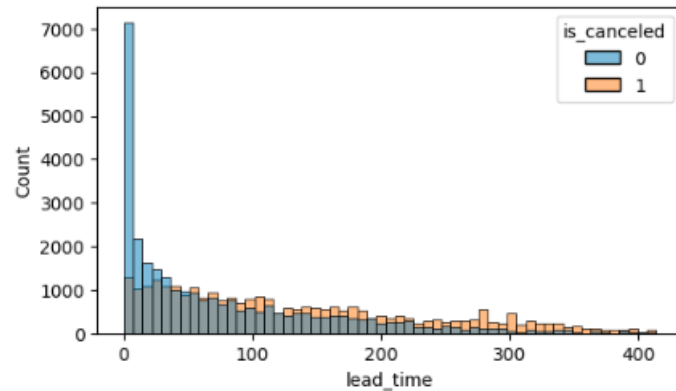
Para imputar los valores de children utilizamos el método de hot deck, buscando valores similares dentro del dataset y elegimos asignar el valor 0.

Para country decidimos eliminar las filas con valores faltantes porque representaban un porcentaje muy pequeño del dataset.

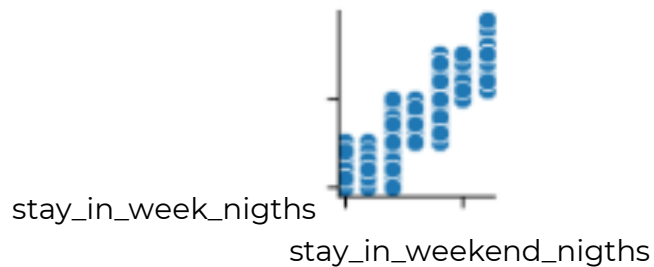
Para agent y company imputamos sus valores con -1 ya que el paper aclara que no se trata de datos faltantes sino de reservas hechas directamente con los establecimientos, por lo que sólo le asignamos este id para poder manipular los datos.

Visualizaciones

El siguiente es un histograma de la variable lead_time y podemos ver que en la mayoría de las observaciones el tiempo de espera es casi nulo. Además nos da información de su relación con el target y podemos ver que la mayoría de las reservas inmediatas (con 0 lead time) no son canceladas.



El próximo gráfico es un extracto del pairplot que muestra la relación entre las variables `stay_in_week_nights` y `stay_in_weekend_nights`. Vemos que estas tienen una relación lineal, lo cual es lógico porque la cantidad de noches durante la semana tendría que condecirse con la cantidad de noches del fin de semana de la estadía.



Tareas Realizadas

Integrante	Tarea
Agustín Ezequiel Sanchez Decouflet	Análisis de Valores Faltantes Detección de outliers univariados Armado de <u>Reporte</u> Análisis de variables cuantitativas Detección de outliers multivariados Visualización de variables
Franco Mazzaro	Análisis de Valores Faltantes Análisis de variables cualitativas Detección de outliers univariados Relaciones entre variables Variables irrelevantes Visualización de variables
Ariadna Antonella Cattaneo	Análisis de Valores Faltantes Detección de outliers univariados Imputación de Datos Armado de Reporte Análisis de variables cuantitativas Relaciones con el target Visualización de variables