Checkpoint 1 - Grupo 29

Integrantes	Padrón
Isidro Héctor Borthaburu	108901
María Delfina Cano Ros Langrehr	109338
Martín Wainwright	108211

Análisis Exploratorio

Primero, el dataset utilizado tenía inicialmente 61913 filas y 31 columnas. Las columnas tenían múltiples variables de diferentes descripciones o características de las distintas reservas de los hoteles.

Las variables más destacadas de este son:

- **Lead time**, que es una variable cuantitativa que muestra el tiempo de antelación con el que se realizó una reserva y nos sirve para analizar si cuánto antes se realizó la reserva, más propenso es a cancelar o no.
- **Arrival date month**, también es un feature importante del dataset y es una variable cualitativa. Esta nos muestra cuál es el mes en el que el/los huésped/es llegan. De modo que puede ser que haya más reservas en los meses de verano en un hotel resort que en un city, por ejemplo.
- Las variables que contiene la cantidad y tipo de huéspedes como **adults, children y babies** puede ser que intervengan a la hora de decidir si cancelar o no. Estas son variables cuantitativas ya que muestran la cantidad que hay en una reserva
- **Previous cancellations** es una variable cuantitativa al igual que **previous bookings not canceled** y podemos observar si es común que el huésped cancele varias veces o no.
- **Is repeated guest** es una variable cuantitativa y muestra si el cliente ya visitó el hotel alguna vez o no. Si este lo hizo, al ya conocer el hotel puede ser que es menos probable que cancele.
- **Deposit type** es una variable cualitativa y muestra si el depósito realizado es retornable o no. Se puede pensar que si el depósito no se puede devolver, el cliente no va a cancelar porque perdería el dinero.

Preprocesamiento de Datos

En la notebook hicimos los siguientes cambios:

1. Columnas eliminadas:
   Company: eliminamos este feature ya que como se demostró en la notebook faltaban el 94.9% de los datos, por lo tanto, sería una especulación muy grande si completaramos las filas vacías con un dato random, o la media o promedio.
2. Correlaciones detectadas:
   En el gráfico 1 del anexo se puede observar cómo las distintas variables se correlacionan entre sí,

- Las variables stays in **weekend nights** y **stays in week nights**, ambas están correlacionadas entre sí, como se observó en el HeatMap. La correlación de Pearson es de valor 0.48870973728040334, como esta es mayor a cero pero no vale uno están relacionadas linealmente pero no de forma perfecta. Los gráficos de dispersión se pueden observar en la notebook
- También previous **bookings not canceled** y **repeated guest** están correlacionadas de manera positiva y su correlación de Pearson vale 0.40602970243434483.
- **Children** y **ADR** están correlacionadas de forma positiva también donde la correlación de Pearson vale 0.35048173774153285 que muestra que están relacionadas de forma positiva
- **Lead time** y **is canceled** también están correlacionadas de forma positiva donde la correlación de Pearson vale 0.2938160884261437
- Por último, **agent** y **company** también están correlacionadas de forma positiva donde el valor de la correlación de Pearson es 0.5149693278016929

3. Columnas recodificadas:

Las columnas que recodificamos fueron: hotel, arrival_date_month, meal, country, market_segment, distribution_channel, reserved_room_type, assigned_room_type, deposit_type, customer_type, id. Estos cambios los hicimos porque para graficar, analizar y realizar acciones sobre algunas features se facilitaba de esta forma

4. Valores atípicos:

Realizamos dos análisis univariados, visual y matemático. El visual es cuando graficamos los Box-Plot y los valores atípicos se destacan porque se encuentran por fuera de los bigotes(máximos y mínimos). Luego de esto, encontramos que las siguientes columnas son algunas de las que más se destacan con outliers: days_in_waiting_list, previous_booking_not_cancelled y ADR. Después el análisis matemático es el caso de z-score donde hay un umbral de tres, podemos ver que las variables que superan esto son stays_in_weekend_nights, adults, days_in_waiting_list, total_special_request, entre otras.

Para el caso del análisis de valores atípicos multivariados también realizamos dos tipos de análisis. Uno de ellos fue "Lof" donde se pueden ver los outliers dependiendo que tan grande es el radio de su circunferencia. En cambio, en el "Mahalanobis" se compara con la densidad de puntos.

Todos los diagramas y análisis completos pueden ser observados en la Notebook, igualmente se puede observar los outliers en el cruce de dos variables en el gráfico 2 del anexo.

Debido a las técnicas que utilizamos, las decisiones que tomamos con respecto a los outliers fueron que los outliers que no eran coherentes y no tenían sentido con el análisis los eliminamos, y luego normalizamos el resto de ellos.

5. Valores faltantes:

Realizando las cuentas correspondientes pudimos observar que las columnas que tenían datos faltantes son **company, agent, country y children**. Por lo tanto, analizamos cada una de estas variables. Primero, la variable **company** tiene un total de 94.9% de datos faltantes. Al ser un volumen tan grande de estos, eliminamos esta columna ya que especulariamos mucho si completamos estos datos faltantes con la moda, promedio o algún valor aleatorio. Luego la variable **agent** tiene un total de 12.7% de datos faltantes. En este caso reemplazamos los datos con la media ya que no se justifica eliminar toda la columna cuando no es ni el 20% los datos faltantes. También analizamos que en la feature **country** solamente faltan 221 datos del total, este es el 4%, entonces completamos con la variable "unknown" en los datos faltantes. De esta forma no especulamos sobre qué valor puede llegar a tener y, si es necesario, en un futuro si tenemos que analizar alguna fila en particular podemos diferenciar entre las que no estaban completadas y los datos ingresados correctamente. Tampoco se justificaba eliminar la columna por tan pocos datos faltantes. Por último, la variable children tiene un muy bajo porcentaje de datos faltantes, en este caso, como son solamente 4 datos faltantes, completamos estos con la moda de la variable donde en este caso ese valor es cero.

Visualizaciones

El primer gráfico es el que mencionamos anteriormente sobre las correlaciones, este es el gráfico 1 del anexo. En cambio, el segundo gráfico muestra cómo es el cruce entre dos variables y en qué casos se cancelo y en cuáles no. Se puede observar que una parte de una media circunferencia está en azul y todo lo de afuera en naranja son los outliers.

Tareas Realizadas

Integrante	Tarea
Isidro Héctor Borthaburu	A, B y C Valores Atípicos
María Delfina Cano Ros Langrehr	A, B y C Informe
Martin Wainwright	A, B y C Valores Atípicos

Anexo:

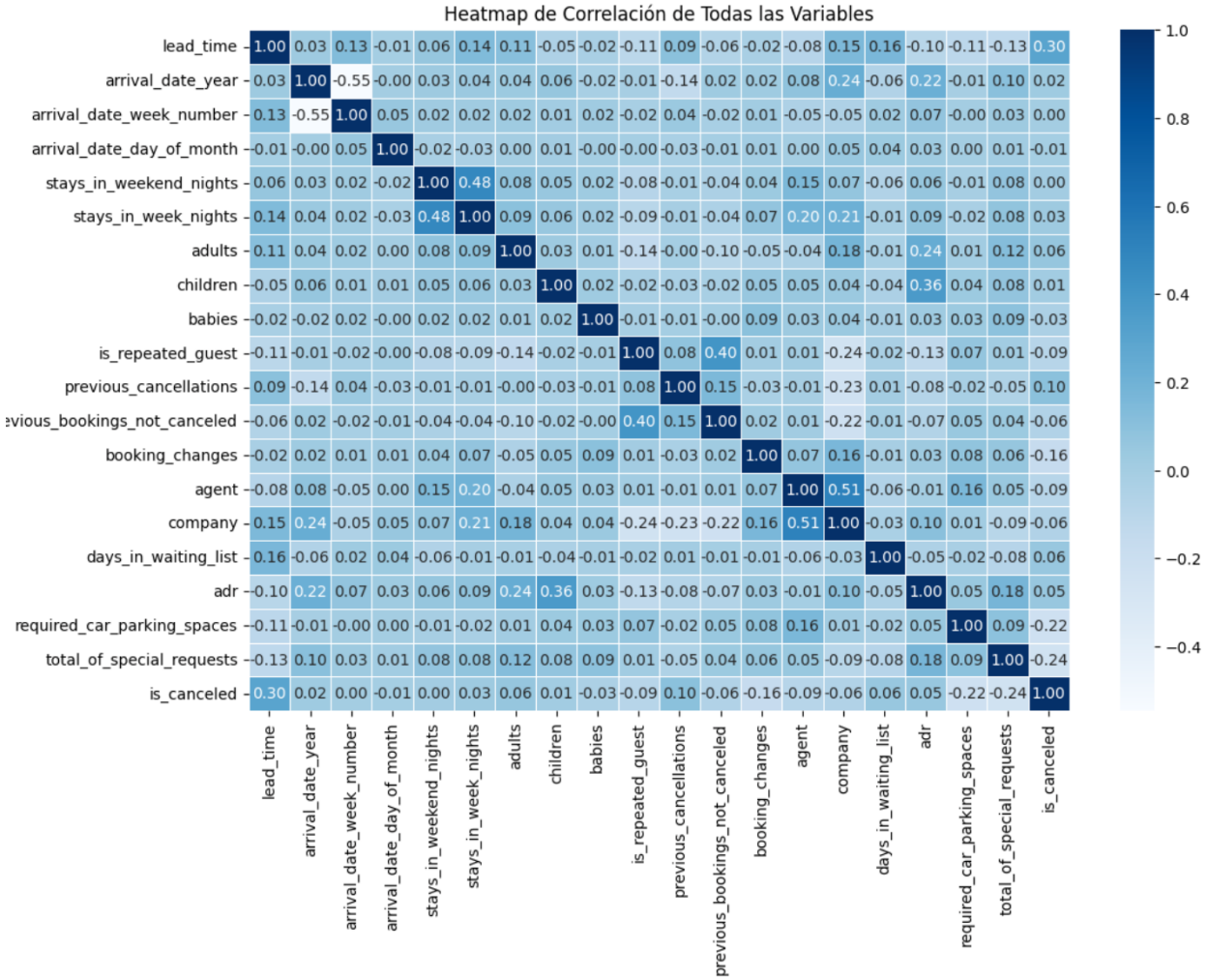

Gráfico 1

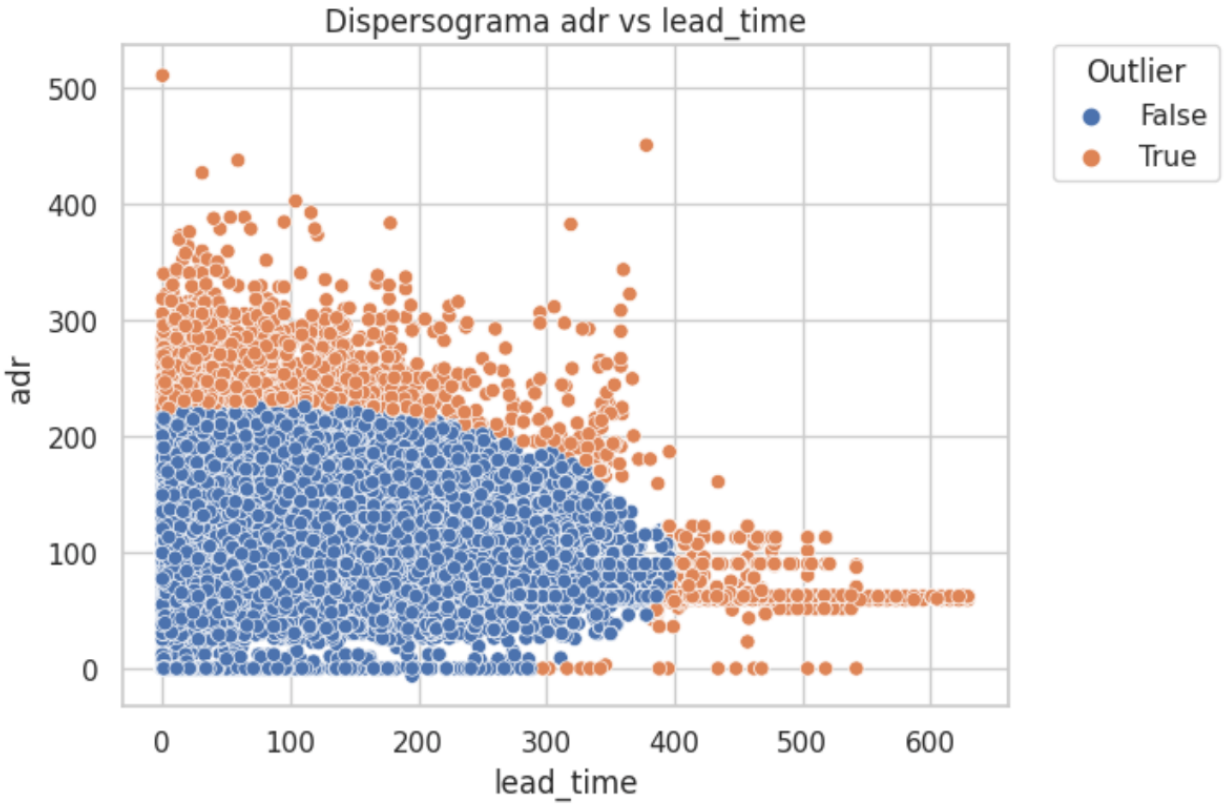
Dispersograma adr vs lead_time
adr
lead_time
0
100
200
300
400
500
600
Outlier
False
True

Gráfico 2