Checkpoint 1 - Grupo 33

Análisis Exploratorio

El dataset analizado contaba con unos 61913 registros, y 31 columnas.
Estaban los nombres de los 2 hoteles de los cuales provenían dichos registros, y entre las demás características que había en estas columnas se encontraban el número de personas que fueron, el tiempo que estuvieron en la lista de espera, la tarifa promedio asociada al tipo de habitación pedida y la cantidad de noches.
También se encuentra la cantidad de veces que estas personas han cancelado otras reservaciones previas, y por supuesto si cancelaron o no la reserva en el registro actual.
Podremos poner esta última en función de las demás características mencionadas, para poder predecir a futuro si alguien podría llegar a cancelar sus reservas en base a la cantidad de veces que lo ha hecho o como cuanto tiempo estuvo en lista de espera o el precio que está pagando, ya que de sufrir modificaciones en la reserva o esperar mucho podríamos pensar que es posible que cancele, por eso resultan tan relevantes estas columnas.
Dichos cambios de reservación se encuentran también numerados en una columna.

Preprocesamiento de Datos

1. Columnas eliminadas:

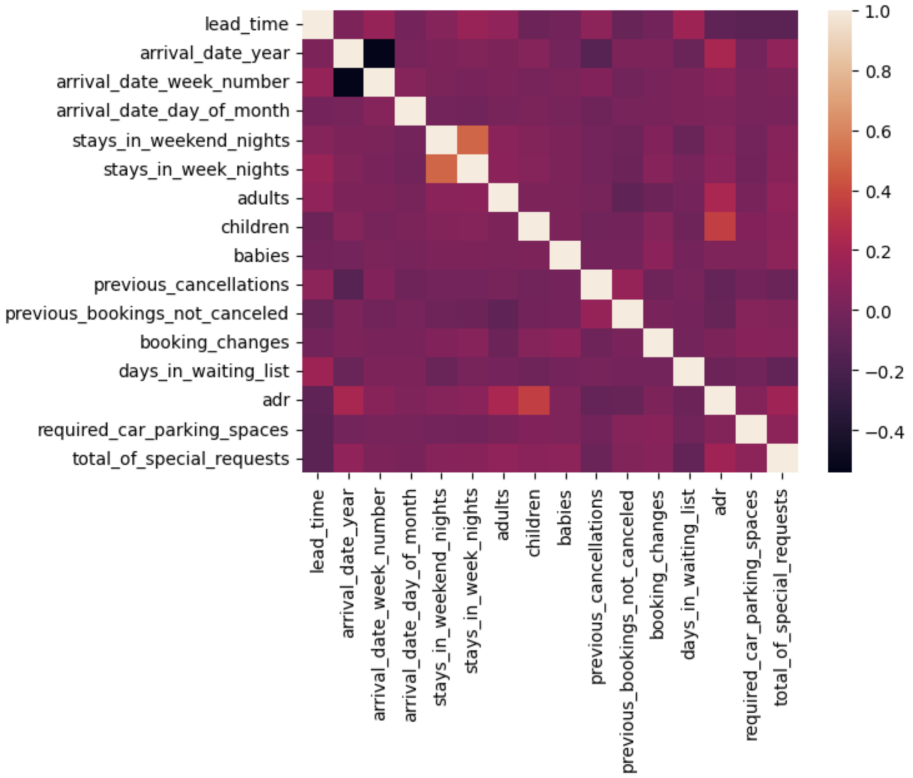
En la columna company teníamos datos faltantes en un 94,9 %, y los pocos que había eran demasiado variados, no había unas compañías en particular de las que se pudiera afirmar que eran las más utilizadas, y debido a esto pensamos que no podríamos rellenar los datos faltantes y que la información que nos da es irrelevante.
También eliminamos la columna ID ya que no queremos entrenar el programa a base de IDs.

2. Correlaciones detectadas:

La columna adr y children tienen un coeficiente de correlación de 0,35, que si bien no representa una relación tan fuertemente lineal, podemos decir que a mayor cantidad de chicos, mayor será la tarifa por día.
Otra correlación bastante presente es la de stays_in_weekend_nights con stays_in_week_nights (correlación de 0.4887), lo cual tiene sentido lógico ya que al quedarte más días de una semana, tiene sentido que además te quedarías los fines de semana para no cortar la estadía por la mitad.

3. Columnas recodificadas:

Decidimos unificar las columnas arrival_date_year, arrival_date_month y arrival_date_day_of_month, que en esencia mostraban la fecha, así que decidimos hacer una columna “date”, que contenga directamente la fecha completa, y no estamos perdiendo información ya que de la misma se pueden extraer por separado estos 3 valores de ser necesario



4. Valores atípicos:

Analizamos de forma univariada cada una de las columnas y encontramos:

- 0 adultos: Este valor no tenía sentido, ya que para hacer una reservación deberías ser mayor de edad. Esto nos hace pensar que fue un error de entrada y los reemplazamos por el valor más común.
- 55 adultos: para valores a partir de 4 adultos el porcentaje de aparición es menor al 0,004% respecto del total, lo cual resulta muy improbable, pero debido a que la clasificación de la reserva es “group”, nos hace pensar que este valor sea menos raro gracias a este análisis multivariado así que decidimos dejarlo como esta.

- 10 children: es solo un caso y fue cancelada la reserva, pero como el resto de los datos no tienen anomalías, decidimos al menos cambiar este valor, y le pusimos el valor más común que es cero.
- 9 babies: solo 8 reservas tienen más de un bebe, de las cuales solo esta tiene más de 2, así que este valor extremo decidimos cambiarlo por el más común que es cero bebés.

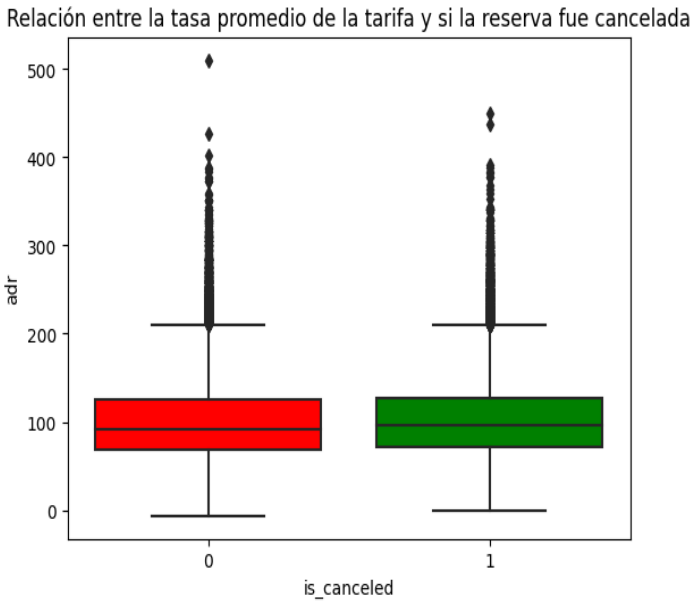
- Adr negativo: es el único caso que presenta esta anomalía, y ni siquiera era un valor cercano a la mediana para pensar que se había puesto el signo por error, así que simplemente lo cambiamos por la mediana.
- Adr 0: hay exactamente 884 casos, que son solo el 0,0147% del total, y decidimos cambiarlos por la mediana.
- Lead time 629: hay varios casos que presentan este valor, entendiendo este como el tiempo que se tarda desde que se hace la reserva hasta la llegada al hotel, si bien es alto, no consideramos que es imposible de ocurrir. Por otro lado, si vemos que mayores a 600 presentan muchas líneas repetidas
- 25 previous cancellations: hay muchos registros duplicados a partir de personas que cancelaron más de 10 veces, y decidimos borrar estos duplicados
- 70 previous bookings not canceled: Hay varios valores tan altos como este, por lo que no los eliminaremos y los consideraremos casos totalmente posibles.
- 17 booking changes: Tenemos muchos valores cercanos por lo cual dejaremos este valor como una posibilidad
- 8 required car parking spaces: acá notamos la gran inconsistencia de que en los registros donde se requiere esta cantidad de espacio de estacionamiento, no hay suficientes adultos, pero al no tener suficiente información del dominio decidimos no modificarlos porque pensamos que quizá exista alguna razón para necesitar esta cantidad de estacionamientos, y son registros que nos podrían servir a futuro.

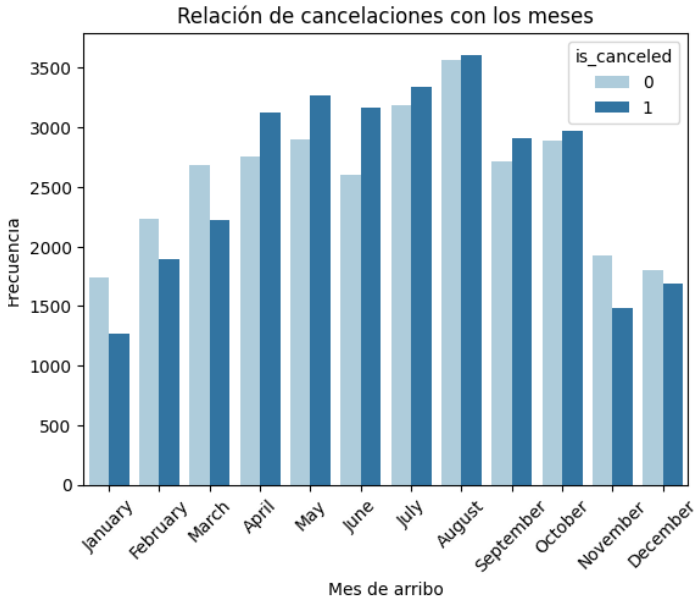
5. Valores faltantes:

En la columna children teníamos solo 4 registros con este dato faltante, por lo que decidimos reemplazarlo con el valor más común que tenemos, que es 0, y no afecta en nada a la distribución.
En agent teníamos un 12,74 % de datos faltantes, decidimos reemplazarlo por el agente más utilizado, porque las proporciones en los porcentajes no cambian de una forma tan significativa, y tampoco había una relación directa con el país, ya que un solo agente salía en muchos paises asi que tampoco se podía rellenar en función al país donde este agente participaba.
En company, debido a que la mayoría de los datos faltaban pensamos que lo mejor sería eliminar esta columna.
En country solo había un 0,36 % de datos faltantes, así que fueron reemplazados por el país más concurrido.

Visualizaciones

Este gráfico es interesante porque muestra que el precio no influye tanto en si uno cancela o no, la media y los cuartiles son muy parecidos.





Tareas Realizadas

Integrante	Tarea
Raimondi Lucas Nahuel	Detección de Outliers Armado de Reporte Análisis de Correlaciones Análisis de Valores Faltantes
Davila Sanchez Manuel	Detección de Outliers Armado de Reporte Análisis de Correlaciones Análisis de Valores Faltantes
Dolores Levi	Detección de Outliers Armado de Reporte Análisis de Correlaciones Análisis de Valores Faltantes