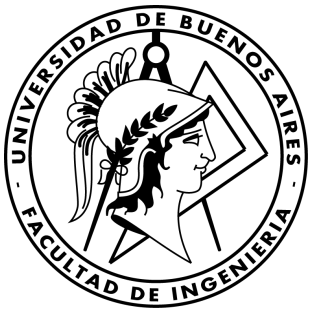


Trabajo Práctico I — Checkpoint I

[75.06/95.58] Organizacion de Datos
Primer cuatrimestre de 2023

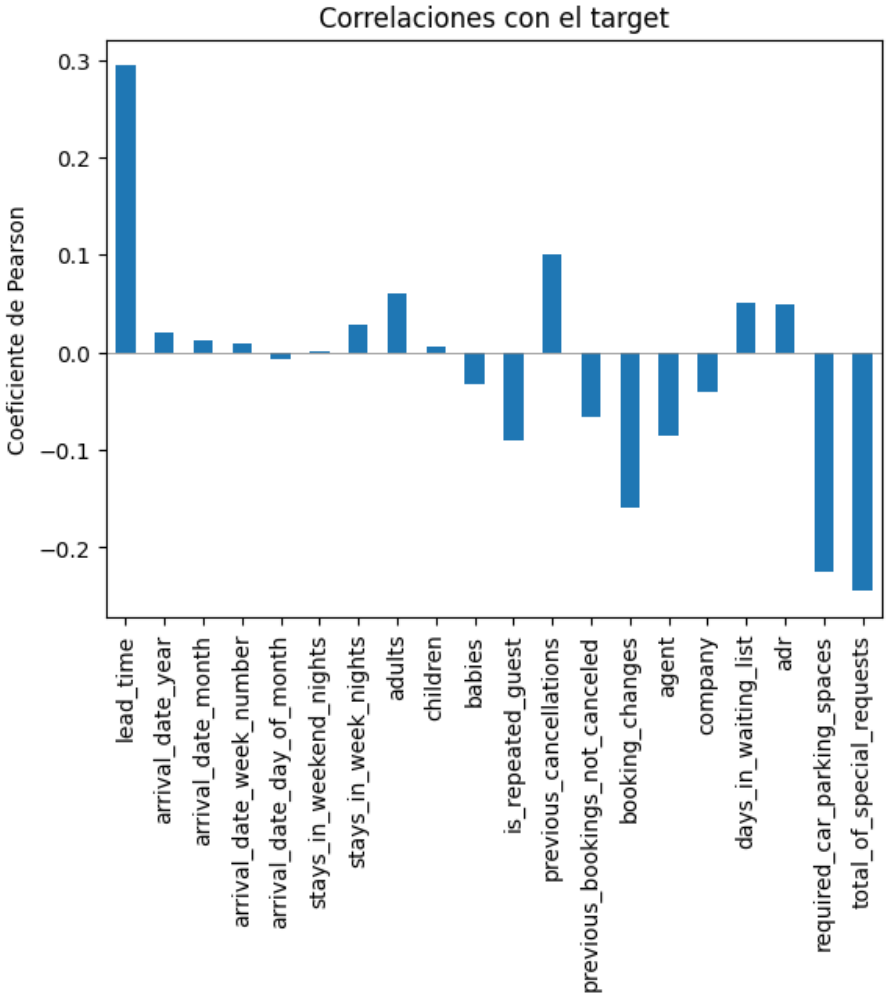
Alumno	Número de padrón	Email
Masri, Noah	108814	noahmasri19@gmail.com
Ayala, Camila	107440	cayala@fi.uba.ar
Loscalzo, Melina	106571	mloscalzo@fi.uba.ar

En este primer checkpoint, el trabajo consistio principalmente en familiarizarse con el dataset. En primer lugar, basandonos en la informacion del paper, explicamos tanto para los lectores externos como para una comprension personal, todas las columnas del dataset. Una vez hecho esto, pudimos clasificar las variables en cualitativas y cuantitativas, para luego explorarlas por separado.

Para las variables cualitativas, calculamos diversas variables estadisticas, como la mediana, la media, la varianza, entre otros, mediante la funcion *describe*. Calculamos aparte la moda. En base a esto notamos ciertos valores atipicos, como un ADR (valor medio por noche) negativo, maximos y minimos, entre otros, los cuales tuvimos en cuenta luego.

Para las variables cualitativas, realizamos un listado de todos los valores que toman, ignorando los valores nulos, y la frecuencia con la que lo hacen. Estos valores, al estar ordenados por cantidad de apariciones, nos permitieron ver facilmente la moda de todas las variables. Notamos que la mayoria contrata mediante una agencia de viajes, que eligen el tipo de habitacion A, entre muchas otras cosas. En la siguiente sección, pudimos ver y analizar esto graficamente.

Comenzamos la etapa de visualizacion calculando y analizando los coeficientes de correlacion de Pearson entre todas las variables. Luego, observamos en particular la correlacion entre todas las variables y el target. Ordenamos los coeficientes en valor absoluto, para ver cuales parecen ser las mas arraigadas, y luego realizamos un grafico de barras de estas.



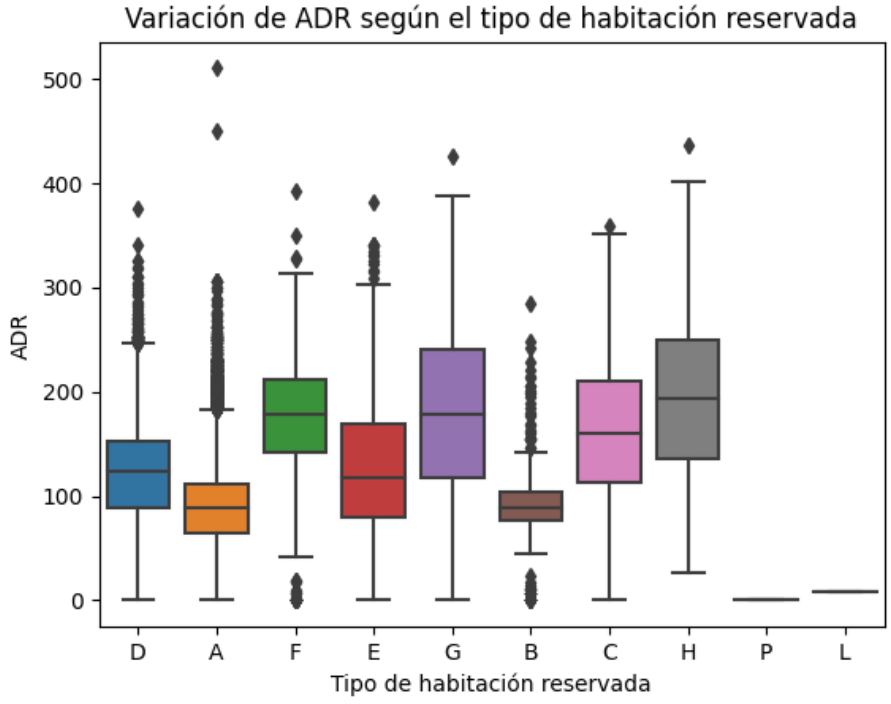
A partir del grafico de calor, comenzamos a visualizar a todas las variables de diversas maneras. En muchos no notamos nada extraño, mientras que en otras logramos encontrar tanto anomalias como cosas que fueron en contra de nuestra intuicion. Entre ellas se destacan:

1. **Reservas sin personas:** tras graficar el adr contra la cantidad de alumnos, notamos que habian reservas sin adultos, lo que luego nos llevo a ver que habian reservas sin personas. Estos valores atipicos decidimos eliminarlos, ya que no representaban un porcentaje muy grande de los datos, y pensamos que ensuciaria mas el dataset modificarlos que lo que perderiamos eliminandolos.

2. **Reservas non refund:** tras graficar el tipo de reserva contra nuestro target, notamos que, mientras que intuitivamente pensamos que poca gente cancelaria una reserva por la que ya pago y cuyo dinero no va a recuperar, mas del 99 % de las personas que contaban con este tipo de reservas las cancelaron.

3. **Reservas en paises anomalos:** tras intentar ver la cantidad de reservas por pais, nos topamos con la existencia de reservas en la Antartida.

   Otra cosa a la que nos dedicamos en esta seccion fue a encontrar el significado del tipo de habitacion. Al graficar la variacion de ADR contra el tipo de habitacion, pudimos ver que la mediana del precio de las habitaciones del tipo A es menor que la de todo el resto.



   En general, el resto aumentan los precios en base al orden alfabetico. Vemos muchos valores por fuera de la caja, y concluimos que podria deberse a la ubicacion del hotel, ya que sabemos que no es lo mismo una reserva en el centro de Nueva York que en las afueras de Houston por ejemplo.

   En las proximas etapas, nos dedicaremos a buscar mas valores atipicos y completar lo que haya quedado pendiente de esta primer etapa y ya comenzaremos con el entrenamiento del modelo.