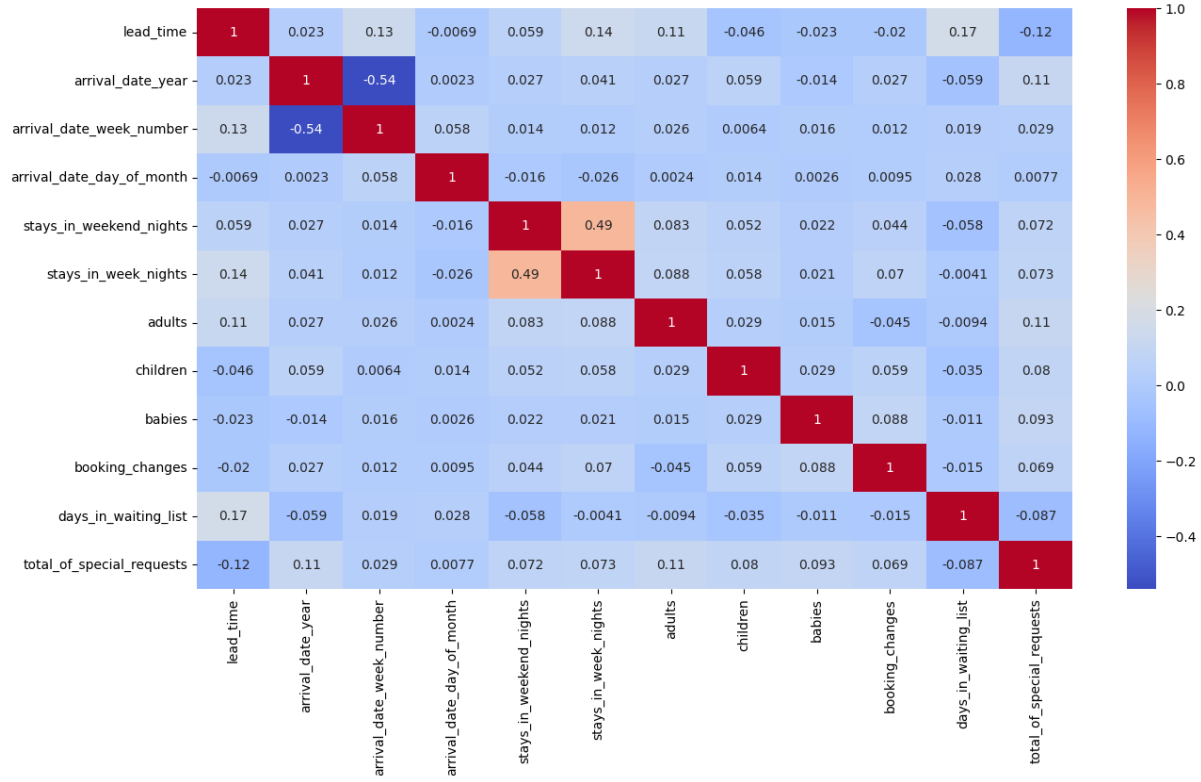
# Checkpoint 1 : Análisis Exploratorio + Ingeniería de Features

En el trabajo presentado se realizó un análisis exploratorio del data set **hotels train.csv**, el cual contiene la información sobre las reservas realizadas a diferentes hoteles a lo largo de varios años.

Primeramente, a partir del paper proporcionado, realizamos un relevamiento de todas las variables que contiene cada dato del dataset, y posteriormente comenzamos a clasificarlas según su tipo: Cuantitativa, cualitativa, y cuasi cuantitativas u ordinales.
En esta instancia surge nuestra primera conclusión: la variable ID es irrelevante ya que contiene tokens que nada tienen que ver con nuestro análisis.

Una vez las tenemos clasificadas, comenzamos a tratar de encontrar y visualizar cómo se distribuyen los datos en esas variables,  haciendo uso de estadísticas descriptivas para las variables cuantitativas (léase media, mediana, moda), análisis de frecuencia para las variables cualitativas, y para ambos casos gráficos de distribución que nos permitan entender los datos de manera más rápida.

Luego procedemos a estudiar cómo se correlacionan nuestras variables cuantitativas entre sí. Esto lo hacemos a través del coeficiente de correlación de Pearson y lo exhibimos mediante el siguiente mapa de calor:




*Heatmap de correlaciones entre variables cuantitativas discretas.*

En el mismo se observa que las variables con mayor correlación son *arrival_date_year* con *arrival_date_week_number*, y *stays_in_week_nights* con

*stays_in_week_nights*. Luego analizamos la relación entre las variables y el target realizando un gráfico de a pares con el target como contraste.

Se realiza también el conteo de casos en que cada variable toma valores nulos y se calcula el porcentaje sobre los casos totales. De este mismo análisis se descubre que la variable *Company* trae alrededor de un 94%. A partir de este análisis es que se decidió eliminar al atributo de *Company* ya que no brinda una información relevante dada su gran cantidad de nulos.

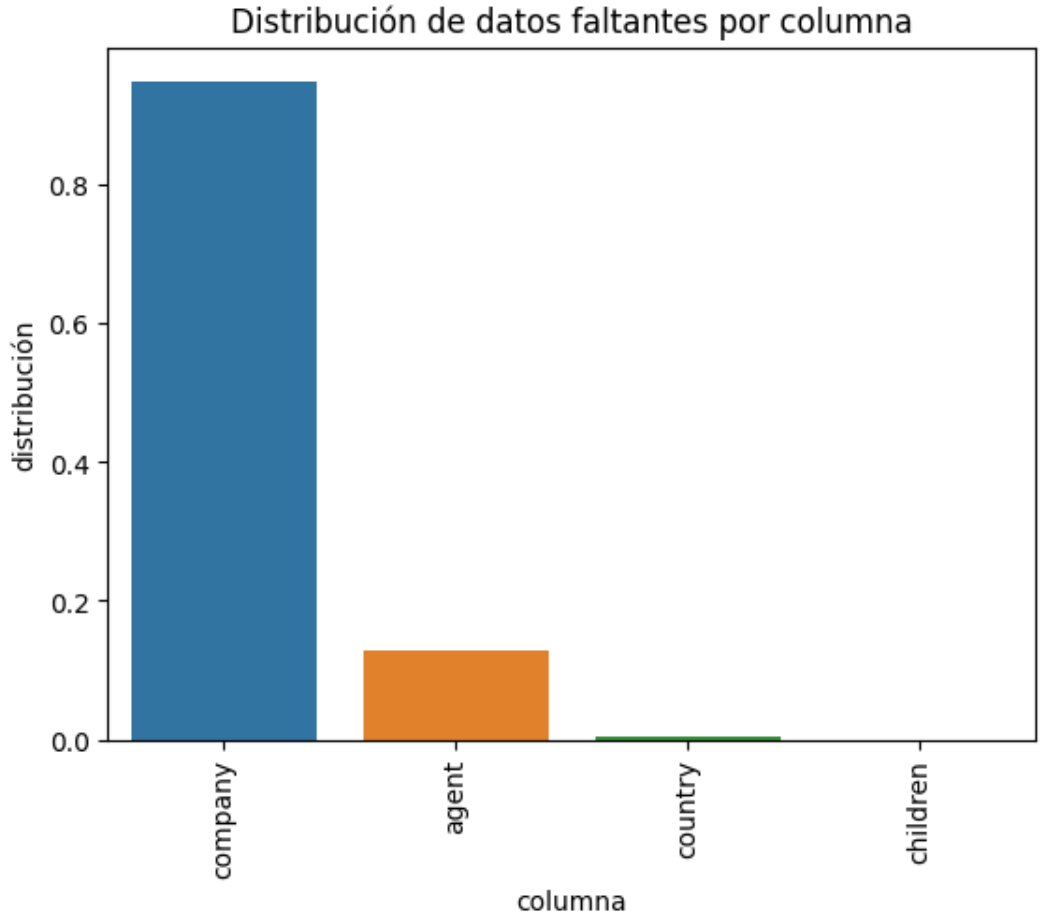



*Gráfico de porcentajes de valores nulos para diferentes variables.*

Luego de la limpieza de datos, con un mejor entendimiento de los mismos, se encontró cierta relación que permitió la creación de una nueva variable. A partir de *reserved_room_type* y *assigned_room_type*, es que generamos la variable misassigned_room, la cual representa con un True o False si la habitación asignada es diferente a la habitación reservada, y de ahí poder sacar alguna relación con el target.