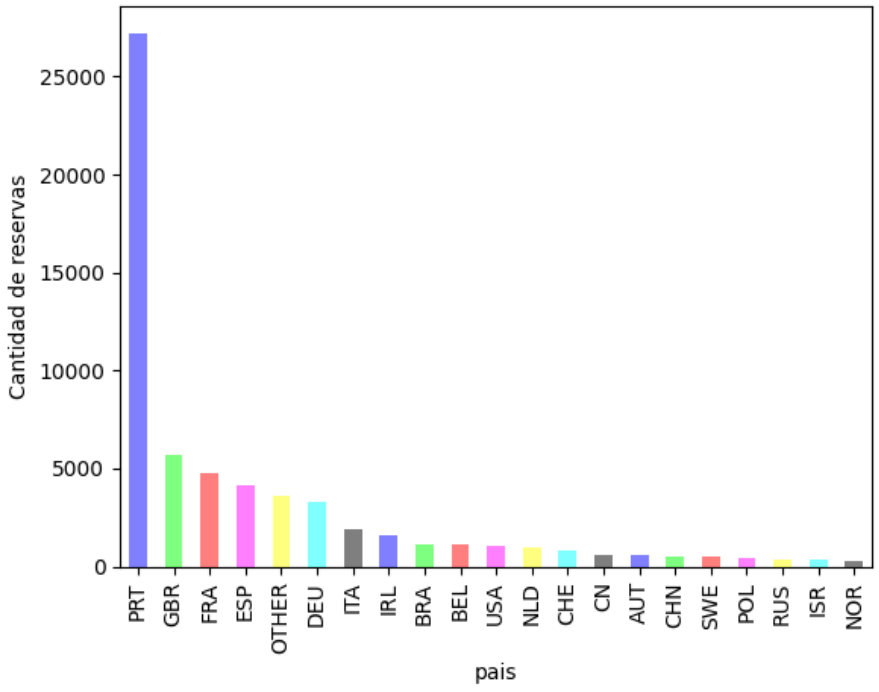
Informe TP1 Reentrega : Reservas de Hotel

Integrantes:
Agama Avila, Arely - 105829
Martinez, Selene Anahi - 100439
Meichtri, Melany - 102330

Checkpoint 1: Análisis Exploratorio y Preprocesamiento de Datos

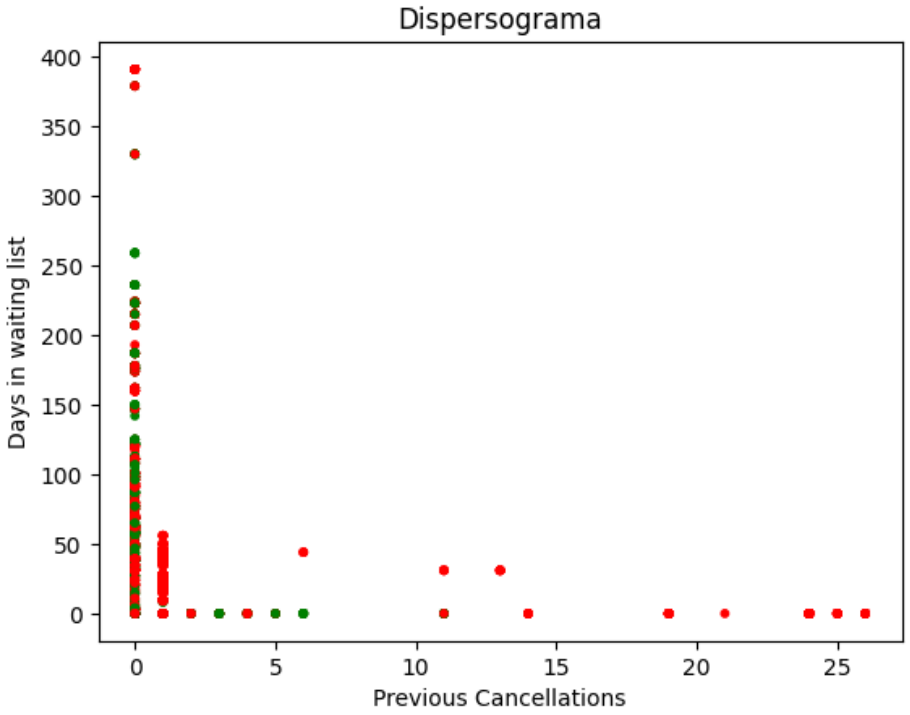
El objetivo es realizar un análisis con el dataset brindado 'hoteles_train.csv', analizando de qué variables dependerá que las reservas sean canceladas. Mostrando gráficos de cada atributo, detectando los datos faltantes y valores atípicos y con ellos tomar una decisión ya sea eliminando o modificando.

a) Exploración Inicial
Básicamente, clasificamos los atributos en variables cualitativas y cuantitativas. Mostrando los valores que toman y la frecuencia en que lo hacen. Decidimos cambiar el nombre del atributo 'adr' dado que no daba mucha información, siendo cambiado por 'average_daily_rate'. Tambien cambiamos 'arrival date month', 'arrival_date_year', 'arrival_date_day', esto para poder juntar estos atributos en uno solo y castearlo a datetime más adelante. Asi mismo, mostramos gráficos de las variables, las correlaciones correspondientes y la relación de las variables con el target(is_canceled).
Debido a la gran cantidad de paises, lo que hicimos fue tomar los primeros 20 con mayor cantidad de reservas y al resto asignarlos en un valor comun llamado "OTHER"

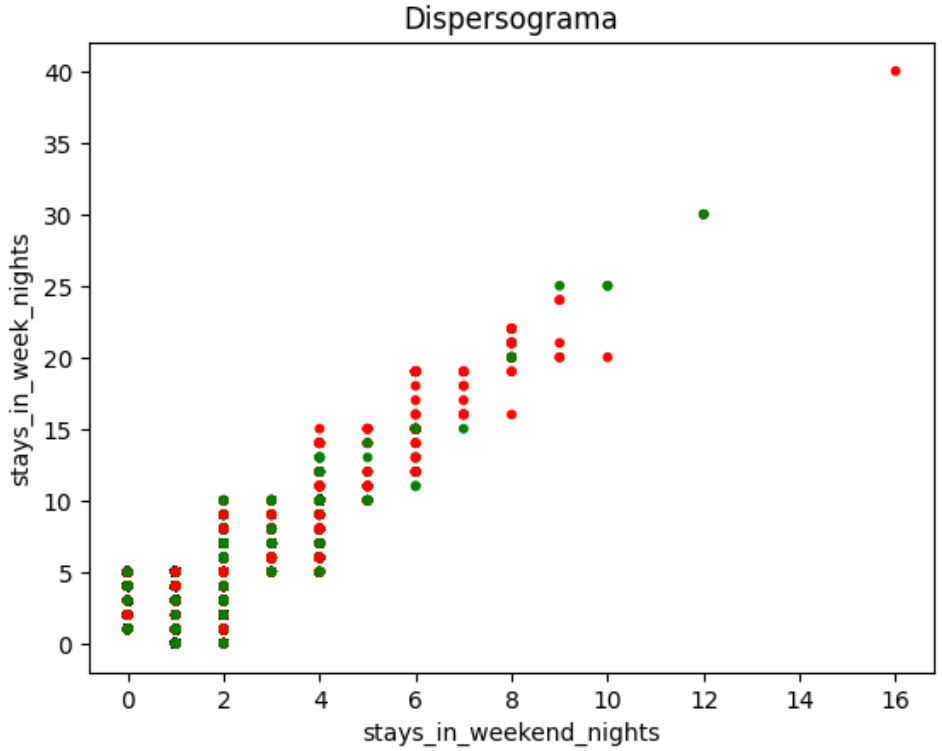


No tuvimos en cuenta reservation_status_date porque reservation_status delataba el resultado de is_canelled

b) Visualización de los datos



Podemos ver que a partir de las 10 cancelaciones previas es mas probable que cancele, al igual que si estuvo mas de 300 dias en waiting list



tambien podemos ver que a partir de las 6 noches en fin de semana y las 10 noches en semana es mas probable que cancele

c) Datos faltantes
Analizamos el porcentaje de datos faltantes encontrados en nuestro dataframe.
Company tiene un 94 % de datos faltantes por lo que tomamos que es una columna irrelevante para el analisis.
En agent decidimos todos los nan ponerlos en un tipo de agent 0.0 ya que este agent no existia previamente

d) Valores atípicos
Analizamos datos atípicos para las variables teniendo en cuenta valores posibles razonables dado el contexto del dataset y donde se concentran la cantidad de valores de las variables.
Datos que datos claramente atipicos son reservas que tengan:

- Solo bebes
- Solo niños con bebes
- Ningun participante
- Sin dias tanto en fines de semana como dias de semana
- Average daily rate en cero
- Average daily rate es negativo
- De grupos de mas de 10 adultos
- De grupos con mas de 8 niños
- Days in waiting list de mas de un año

Decisiones que tomamos:

- Eliminar adr menor a 0 ya que es una sola fila
- Eliminar adr igual a 0 ya que representa el 1.42%, en el contexto es extraña una reserva sin gasto diario
- Eliminar reservas que tengan solo bebes y niños o solo bebes
- Eliminar reservas sin participantes ya que representan el 0.020% del dataset
- Eliminar reservas sin dias en la semana y sin dias en fin de semana
- Dejar grupos de solo niños
- Dejar grupos de mas de 10 adultos y de mas de 8 niños ya que si bien son bien son datos atipicos son datos posibles dados en el contexto