

Análisis Reservas Hotel

Predicción
CANCELACIONES

SUPERVISADO

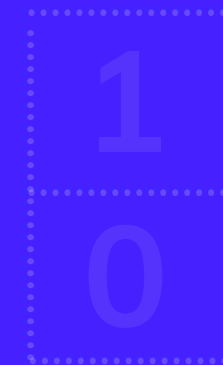
1
0

LAURA LEDO

Objetivo

Conseguir un modelo de predicción para saber si van a cancelar una reserva del hotel.

SUPERVISADO



Dataset Kaggle

- Datos del 1 julio 2015 al 31 agosto 2017
- 119390 reservas
- 32 features
- 2 hoteles

2 HOTELES

40060 RESERVAS



Resort

Hotel situado en la zona del Algarve

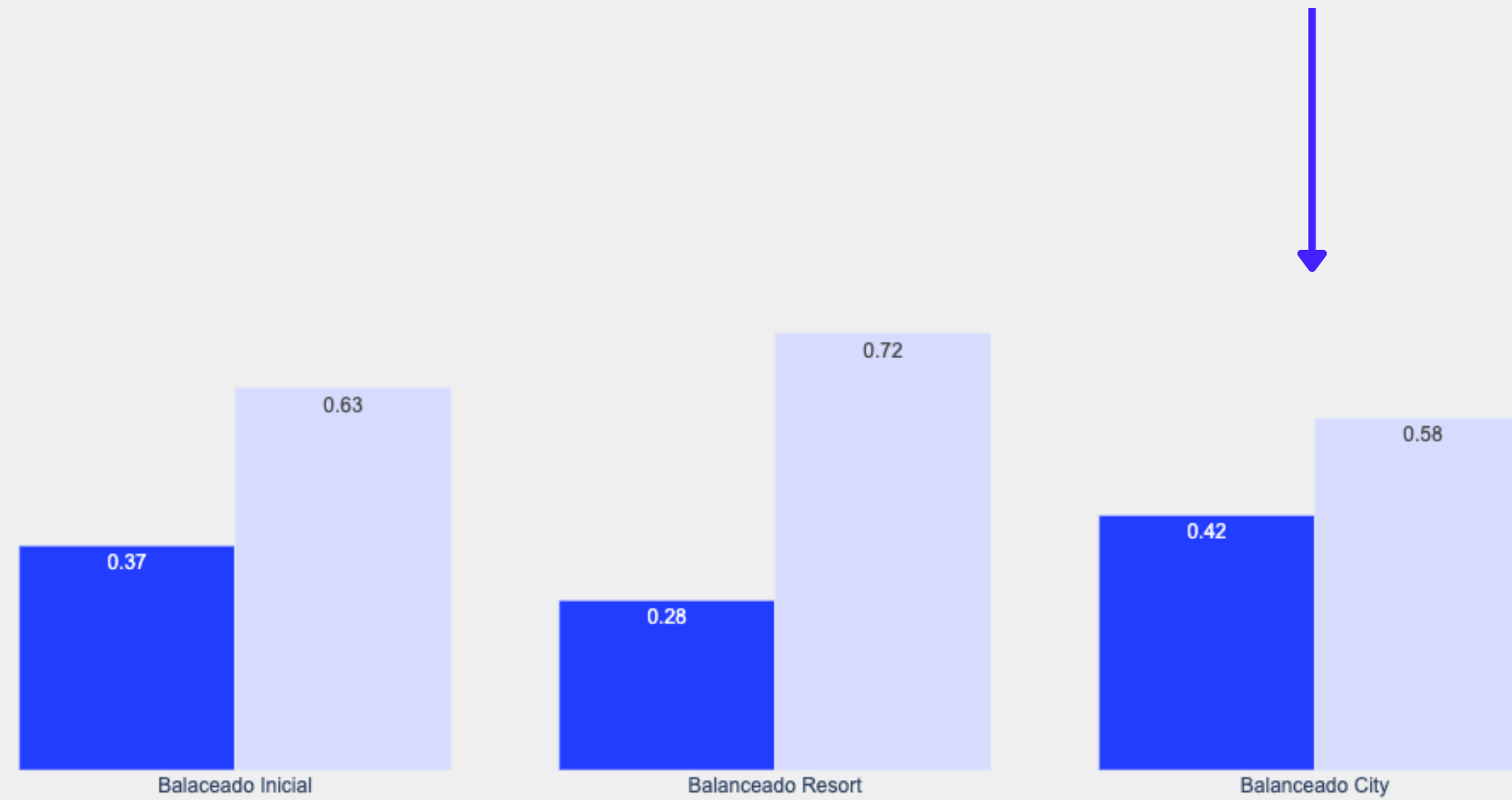
79330 RESERVAS



Ciudad

Hotel situado en la ciudad de Lisboa

Target

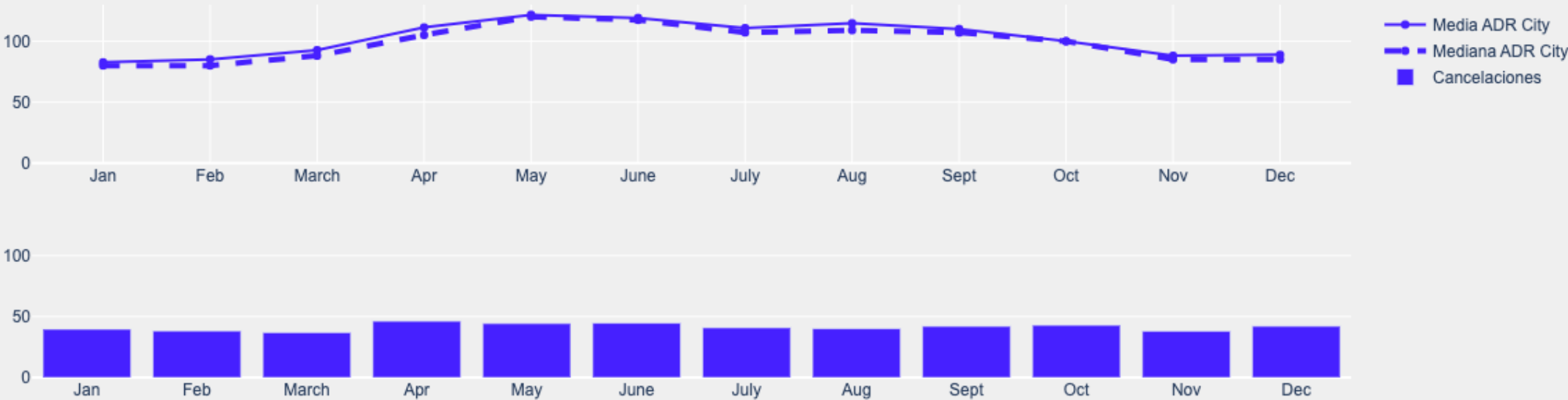


ADR

.....

Tarifa Media Diaria por habitación ocupada:

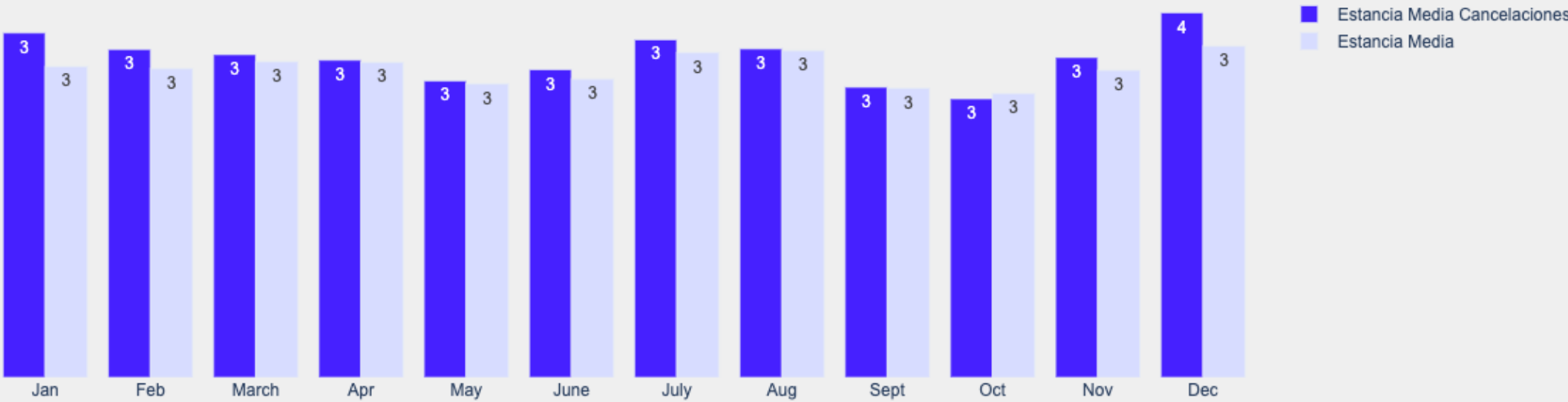
80-125€



Estancia Media

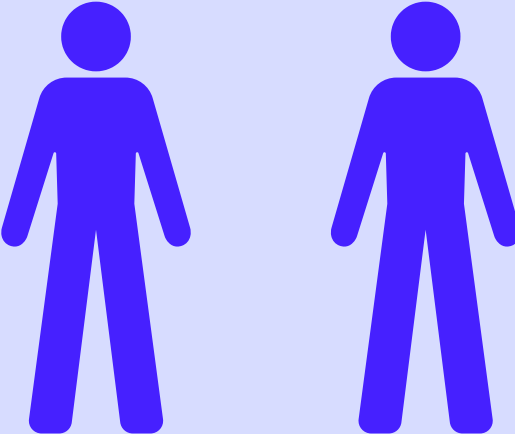
.....

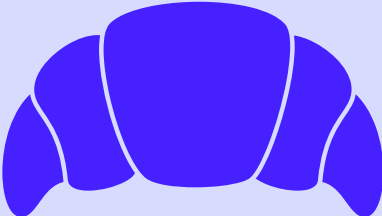
3 noches

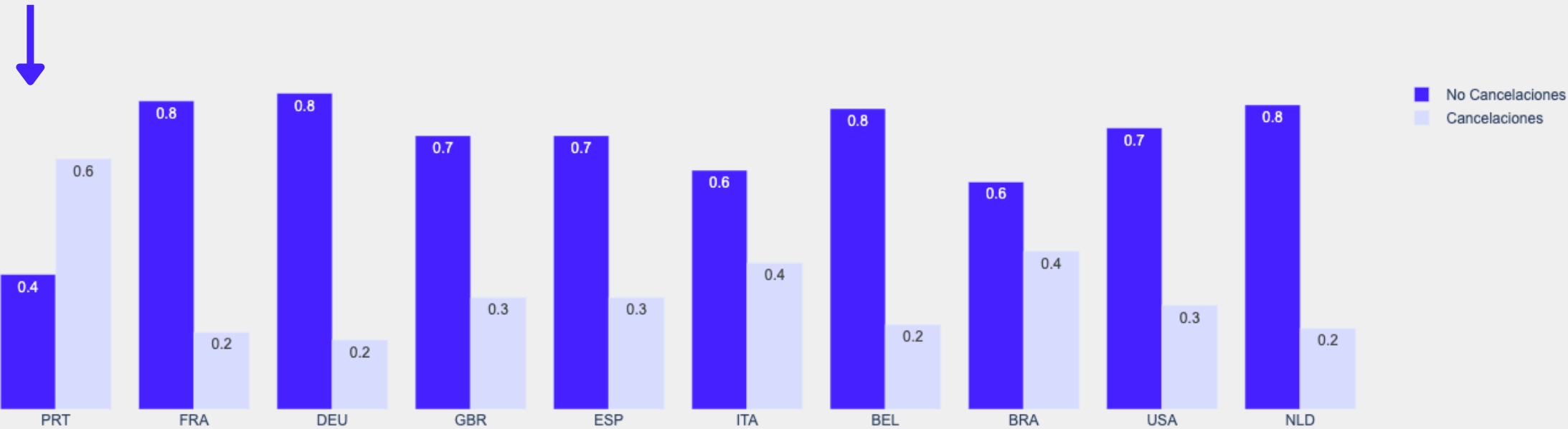


Perfil Huésped

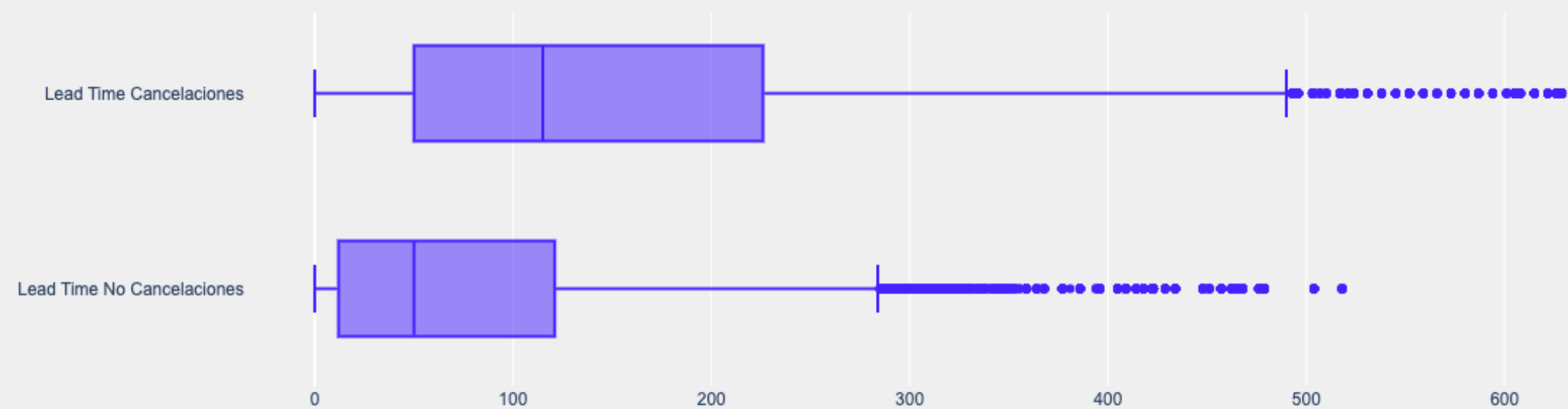
2 ADULTOS




MEDIA
PENSIÓN



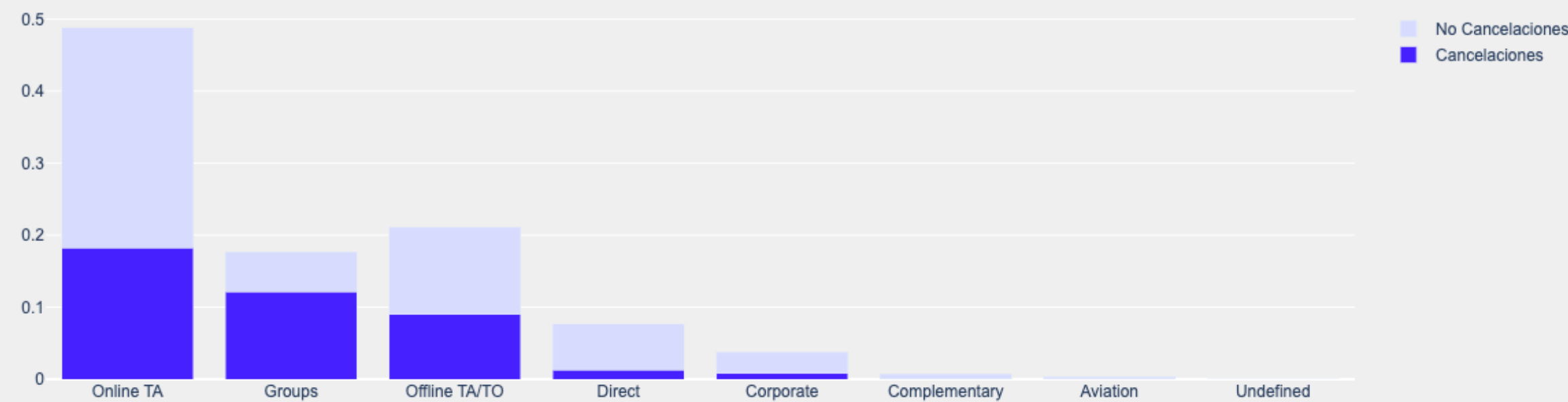
Reservas




Las cancelaciones suelen reservarse con más antelación:

12 - 121 días

50 - 226 días



A large, light purple number '01' is positioned on the left side of the slide, partially cut off by the edge. The '0' is a simple, rounded shape, and the '1' is a vertical bar with a small diagonal stroke at the top.

VARIABLES

15

Variables numéricas

adr

lead_time

days_in_waiting_list

dias_modificacion_llegada

previous_bookings_canceled

previous_cancellations

booking_changes

total_of_special_requests

required_parking_spaces

stay_in_weekend_nights

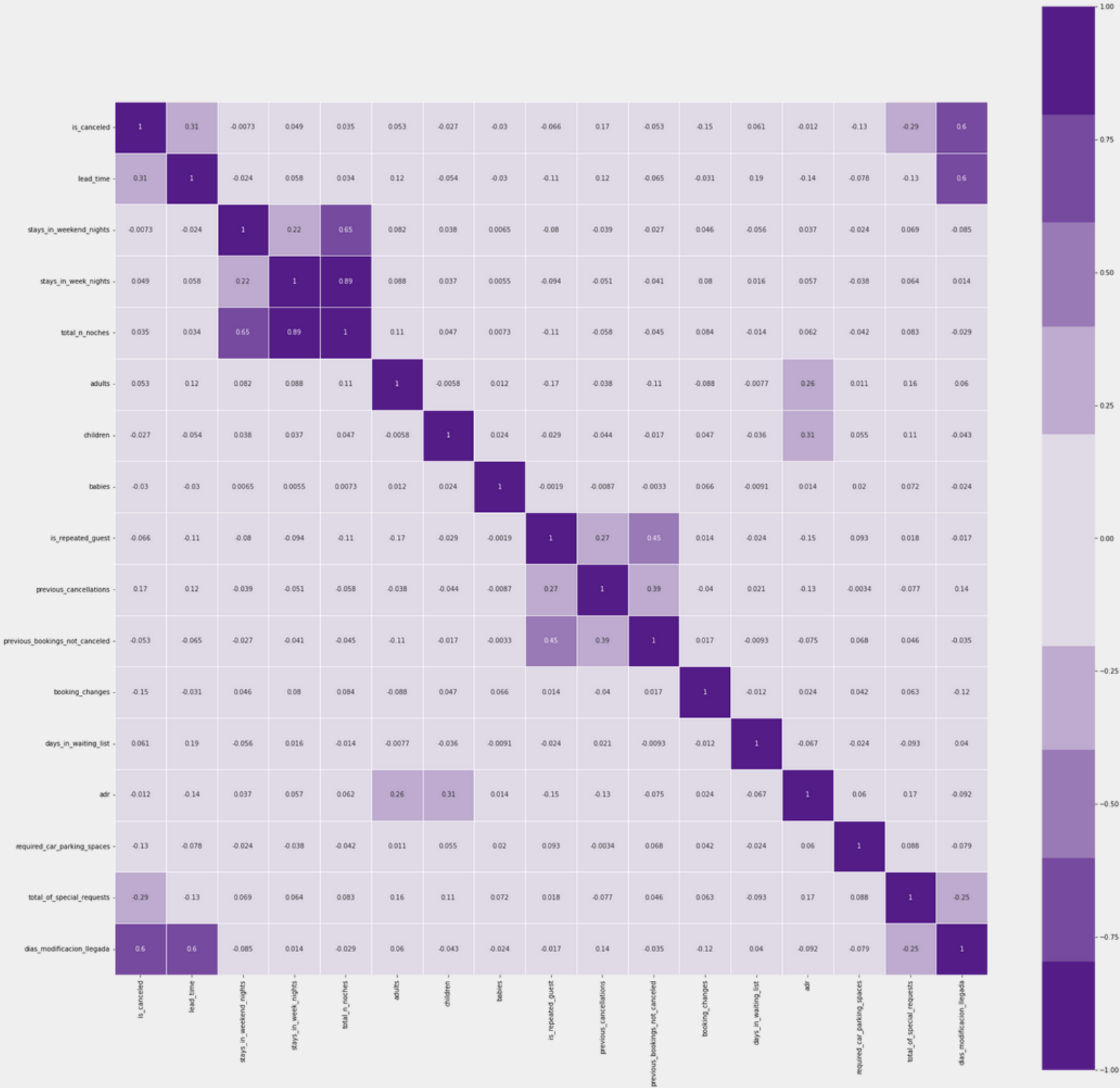
stay_in_week_nights

total_n_noches

adults

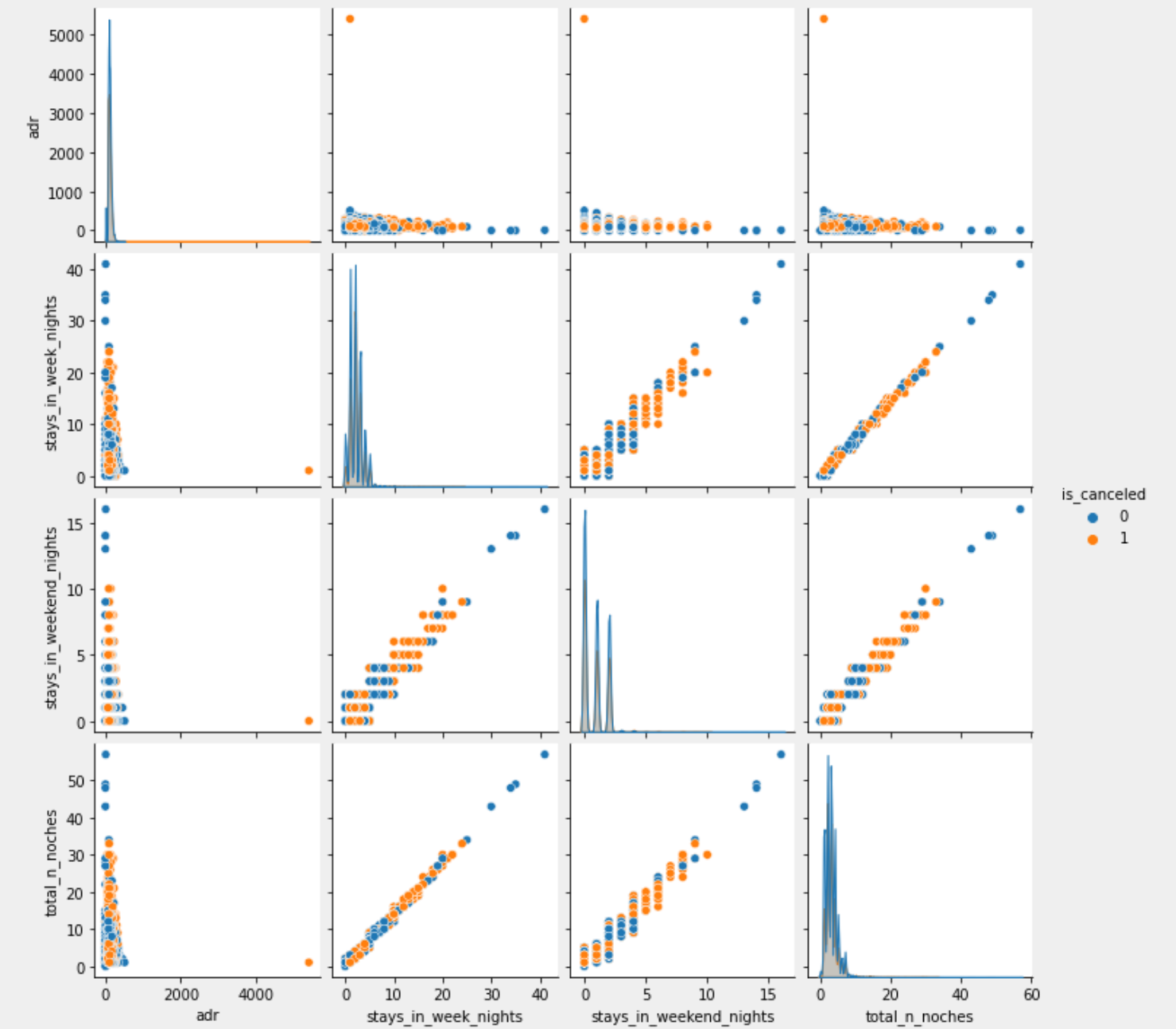
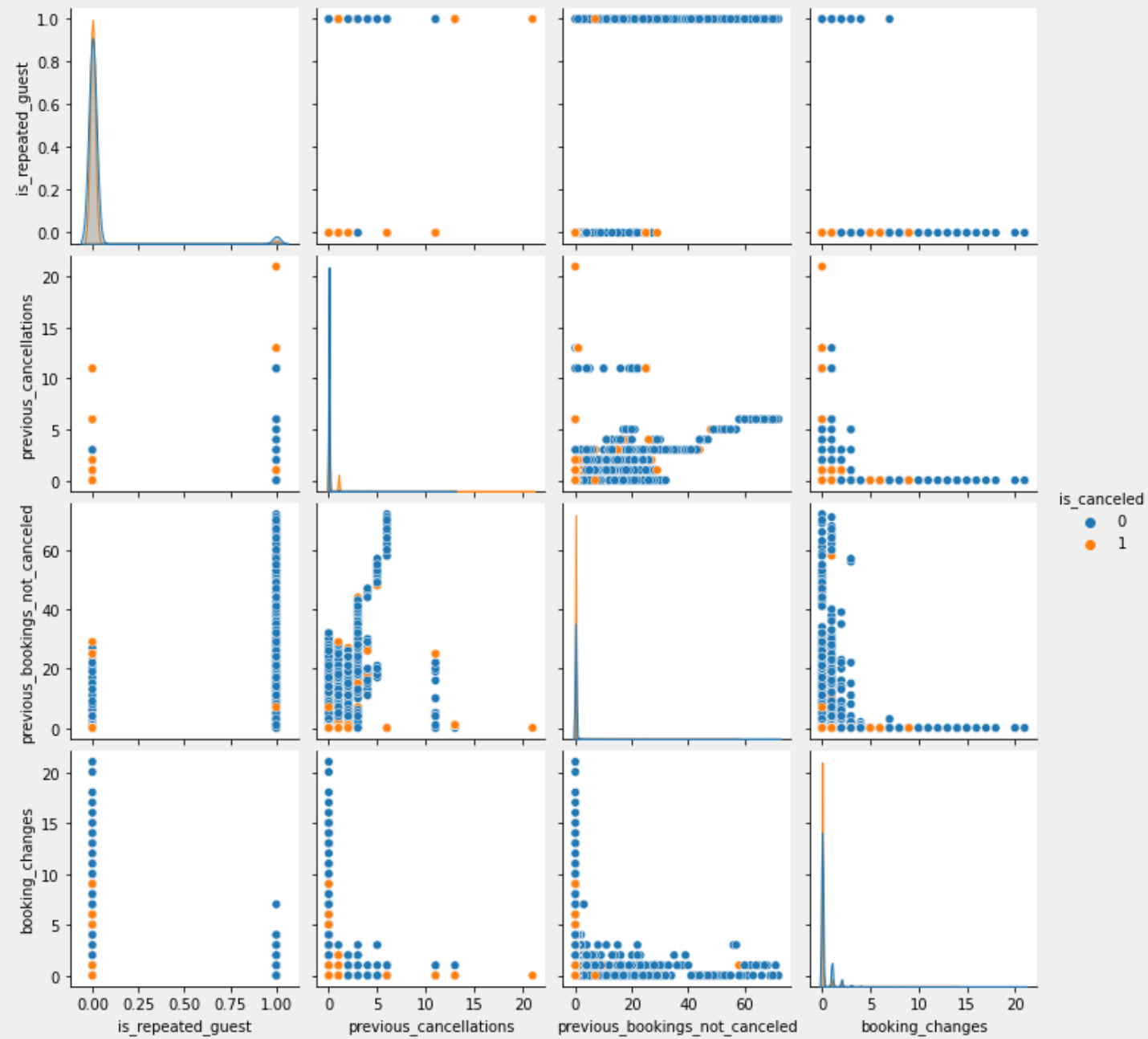
children

babies

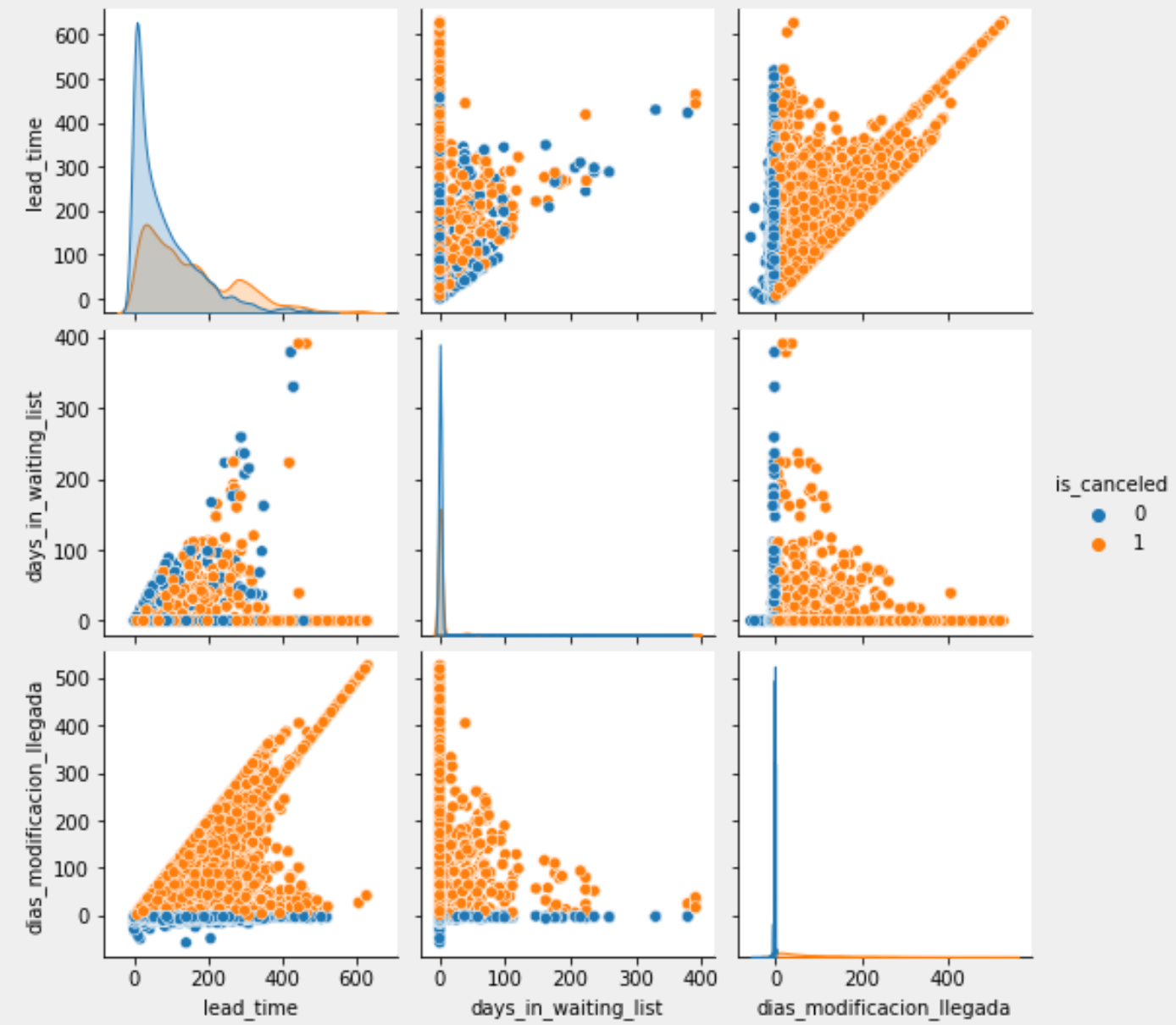


- No incluiremos en el modelo predictivo las variables con fondo azul claro por correlación con otras.

NO TENEMOS CORRELACIÓN LINEAL CON EL TARGET

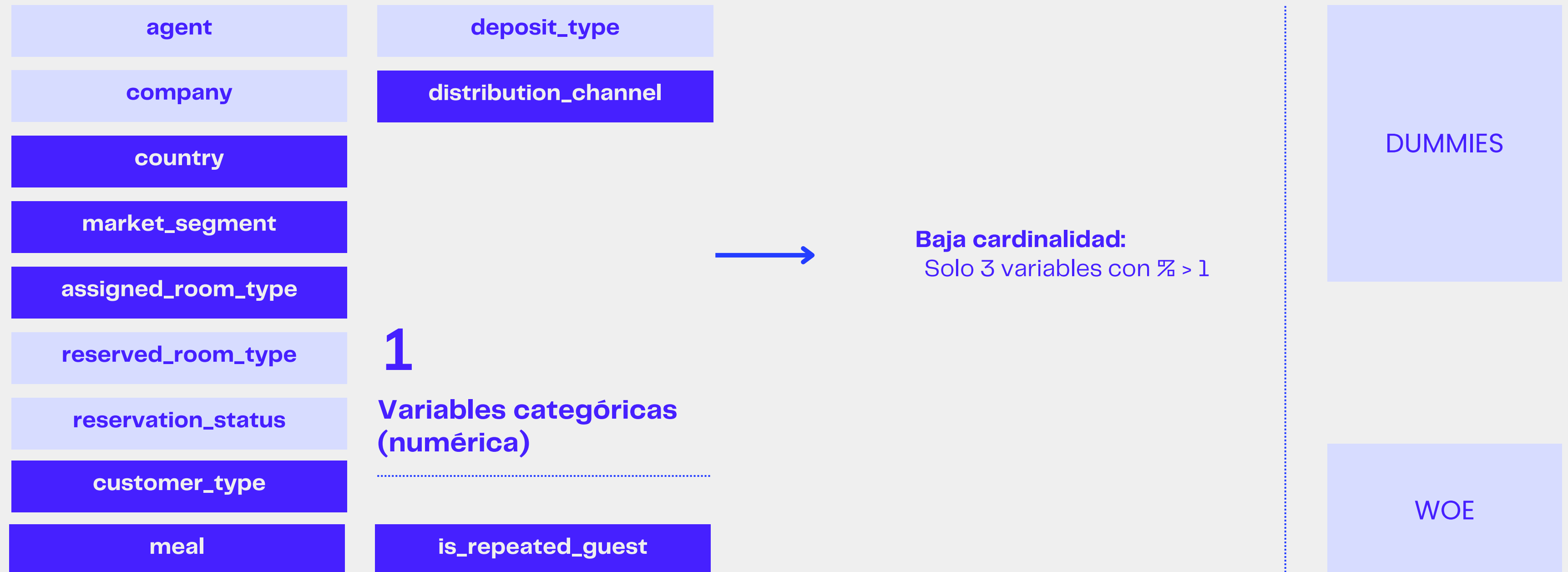


NO TENEMOS CORRELACIÓN LINEAL CON EL TARGET



12

Variables categóricas



- No incluiremos en el modelo predictivo las variables con fondo azul claro por correlación con otras o por demasiada relación con el target (contaminan el modelo).

6

Variables fechas

reservation_status_date

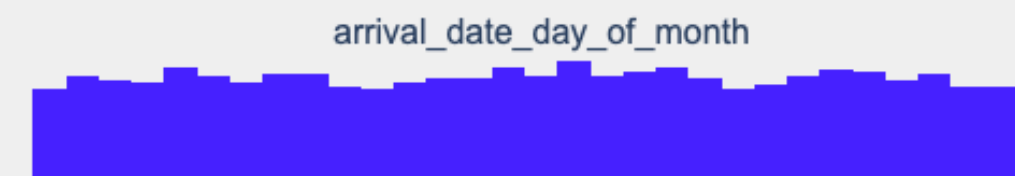
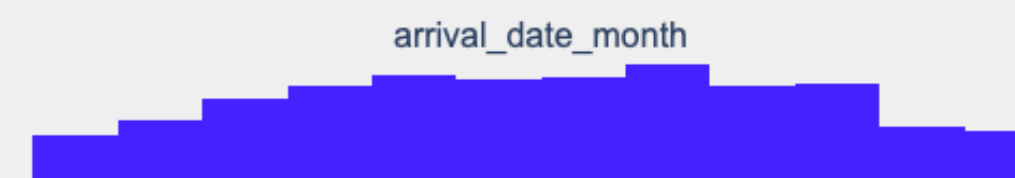
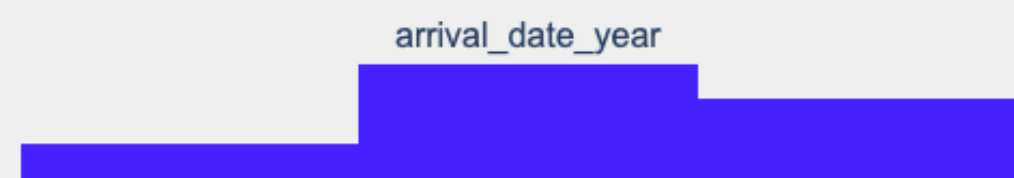
arrival_date

arrival_date_month

arrival_date_week_number

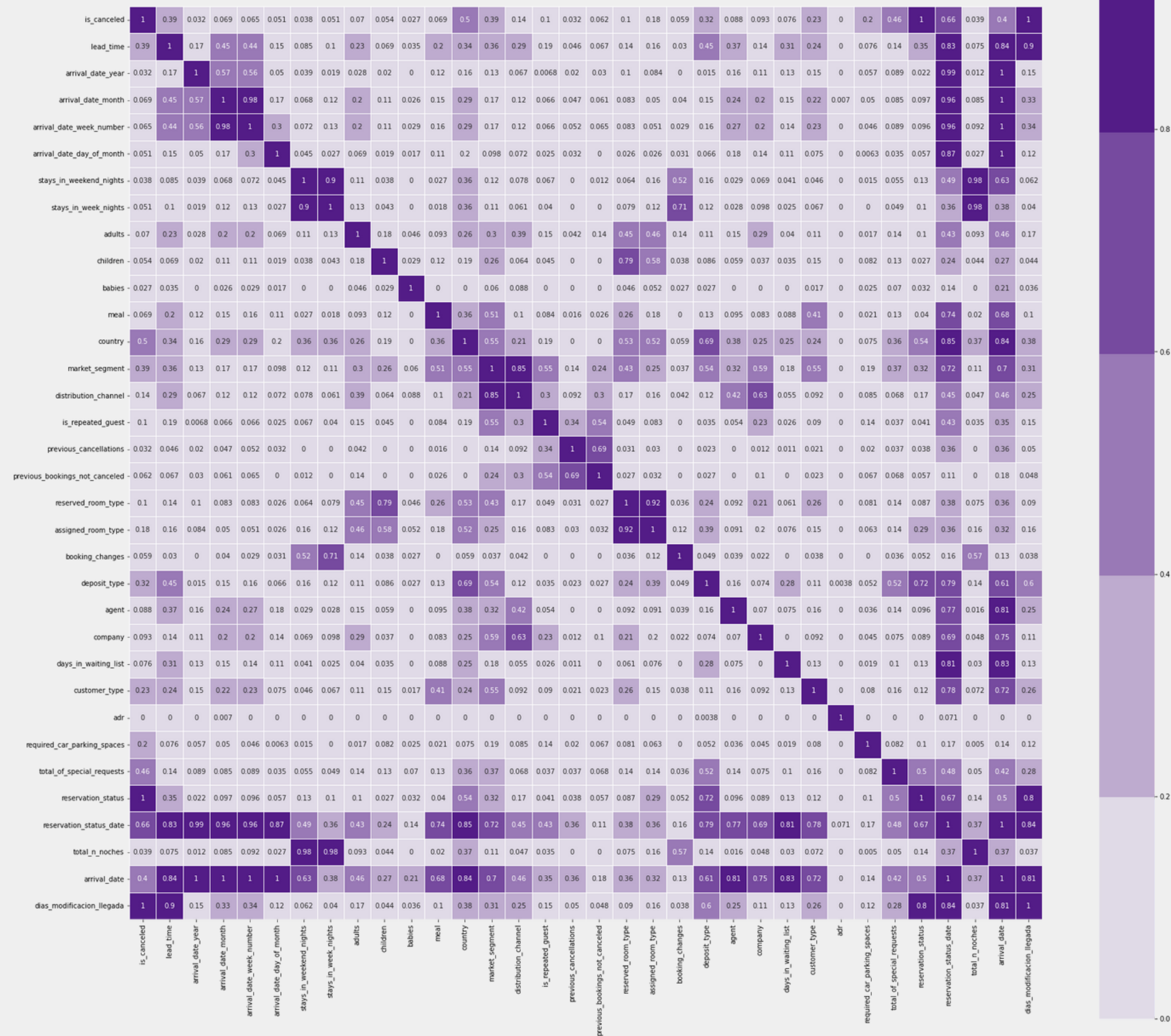
arrival_date_year

arrival_date_day_of_month



- No incluiremos en el modelo predictivo las variables con fondo azul claro por correlación con otras.

MATRIZ PHIK
NO HAY CORRELACIÓN LINEAL

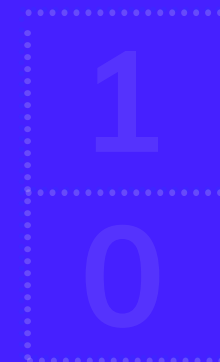




MODELO PREDICTIVO

PROBLEMA DE CLASIFICACIÓN BINARIA

SUPERVISADO



- Random Forest
- SVM
- Bagging Random Forest
- Gradient Boosting Classifier
- XGBoost
- Red Neuronal

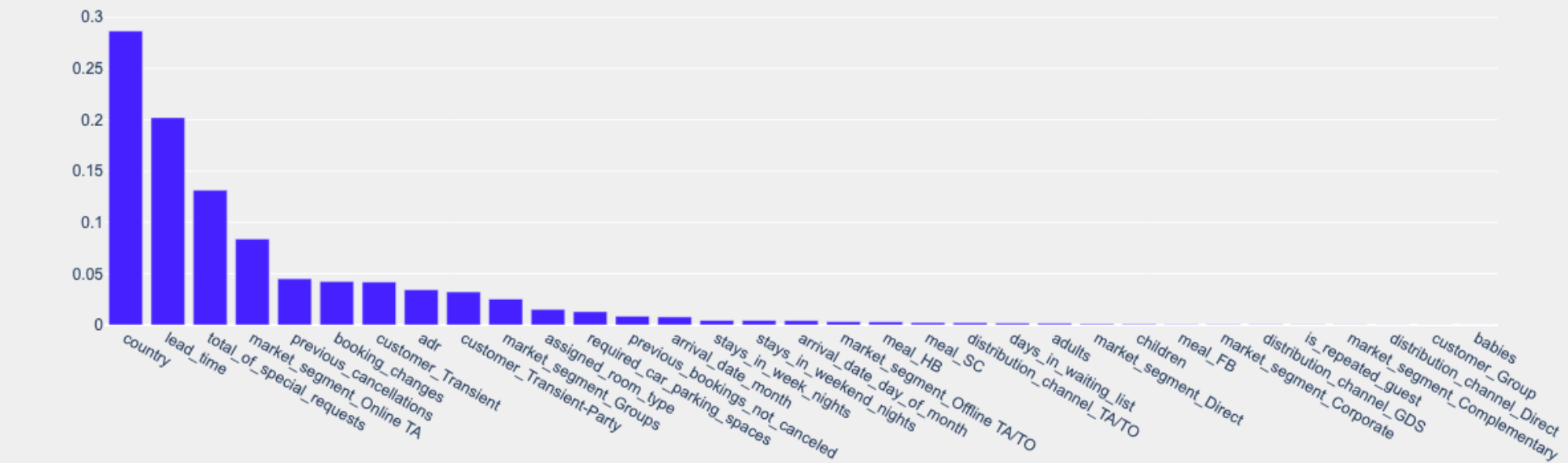
Modelo		Precisión		F1 Score		Accuracy		AUC	
		Train	Test	Train	Test	Train	Test	Train	Test
Gradient Boosting Classif.		0.84	0.84	0.81	0.81	0.84	0.84	0.84	0.83
SVM <small>(con feature selection)</small>		0.85	0.84	0.78	0.78	0.81	0.81	0.82	0.82
Bagging Classifier Tree		0.83	0.83	0.74	0.74	0.80	0.80	0.78	0.79
MLC (Red Neuronal)		0.84	0.81	0.85	0.83	0.88	0.86	0.88	0.85
UN POCO DE OVERFITTING	Random Forest	0.93	0.87	0.90	0.84	0.91	0.87	0.91	0.84
	XGBoost	0.99	0.87	0.99	0.85	0.99	0.87	0.99	0.87

Pasos a Mejorar

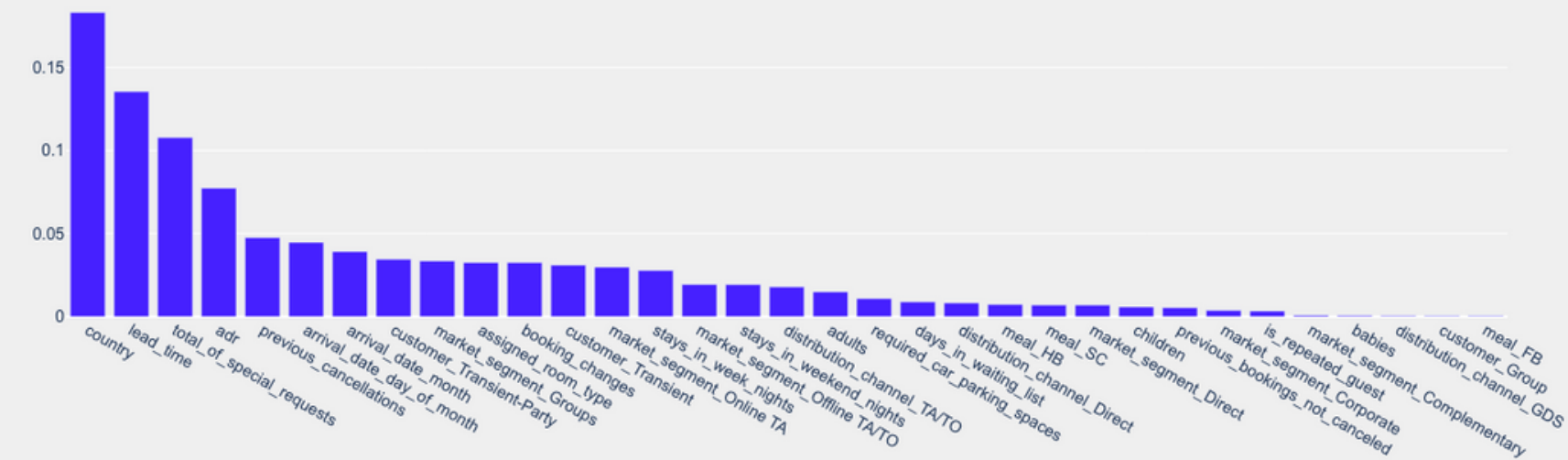
- Probar con otras opciones de parámetros para intentar aumentar el scoring
- Encontrar posibles variables transformadas para mejorar la predicción

FEATURE IMPORTANCE

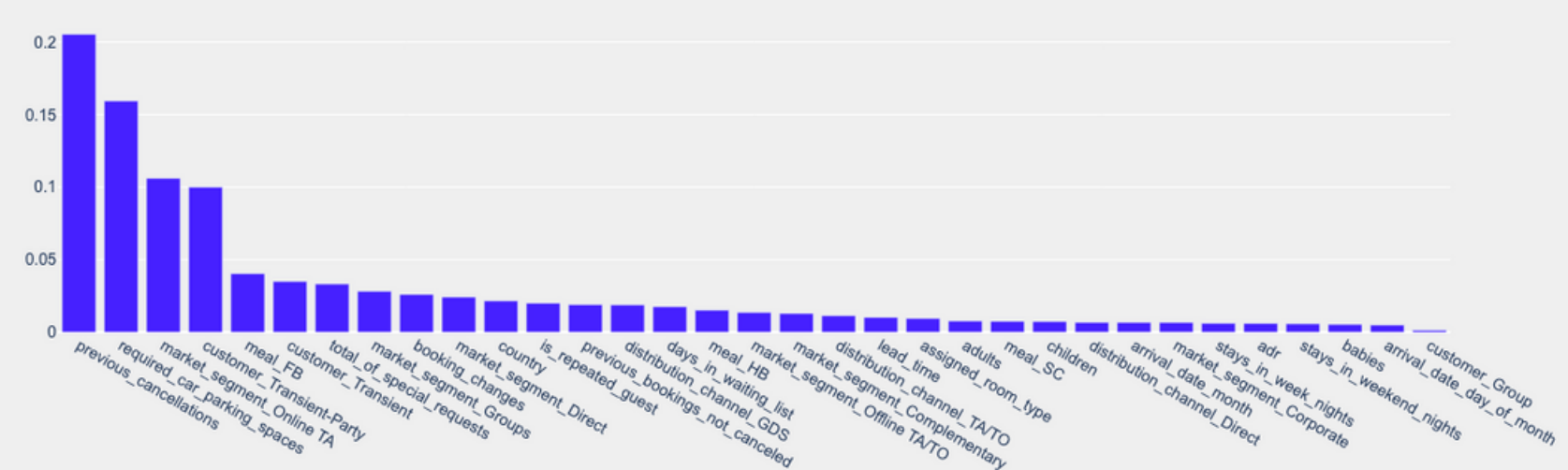
Gradient Boosting Classifier



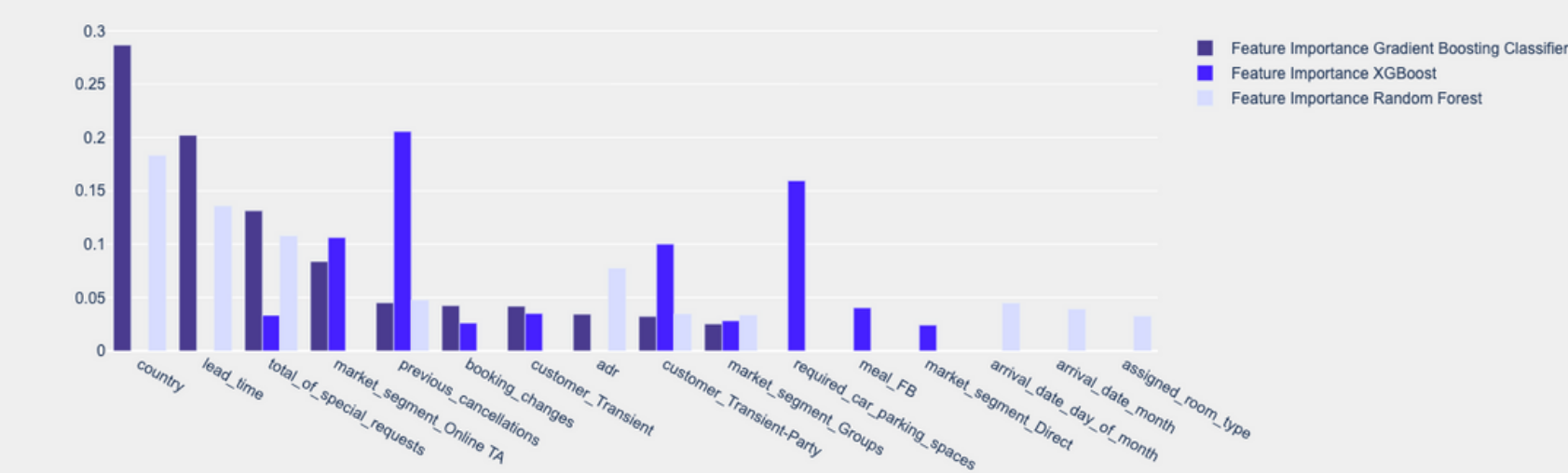
Random Forest



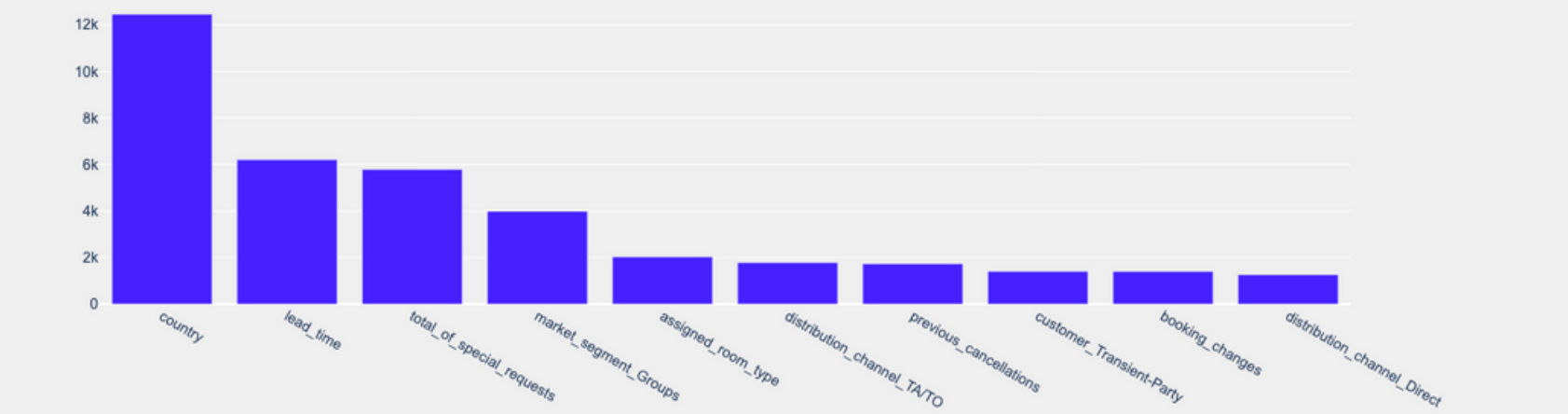
XGBoost



TOP 10 RF, GB y XGBoost



Select KBest



VARIABLES MÁS IMPORTANTES

- country
- lead_time
- arrival date day of month
- arrival date month
- total of special requests
- required car parking spaces
- market segment
- previous cancellations
- booking changes
- customer (Transient)

Variables Feature Importance

country

Los huéspedes procedentes de ciertos países tienen más probabilidad de cancelar:

PORTUGAL : 60% de las cancelaciones
ITALIA: 40%
BRASIL: 40%

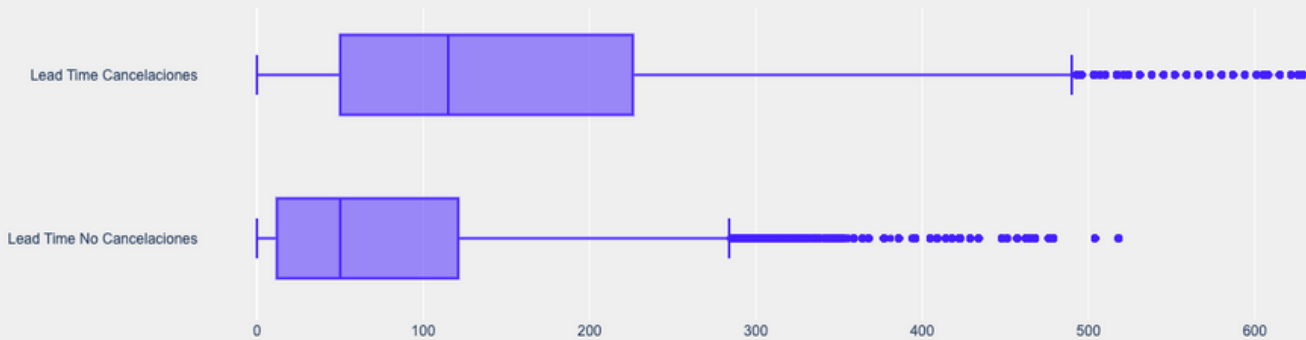
tienen más probabilidad de cancelar.



lead time

Tiempo entre la fecha de reserva y la fecha de llegada.

Como vimos al principio, cuánta mayor antelación, mayor posibilidad de cancelación.



Variables Feature Importance



total of special request

required car parking spaces

booking changes

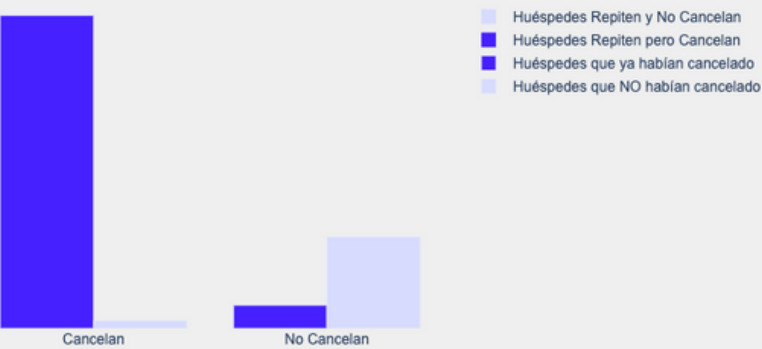
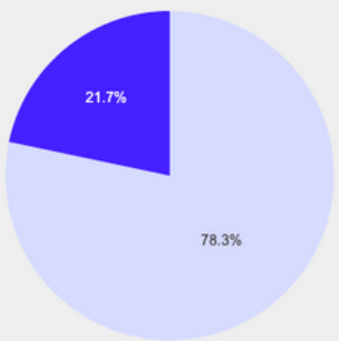
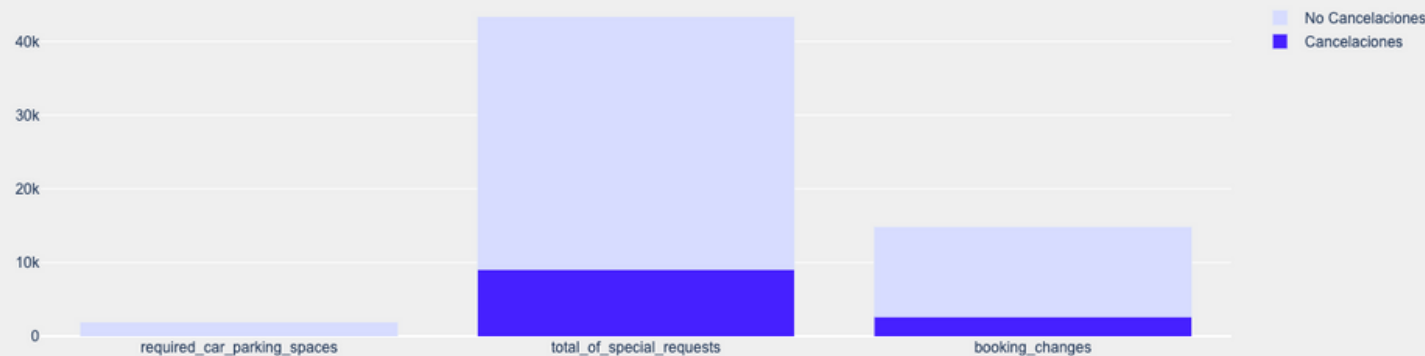
Lo más común en estas columnas es 0. Que algún huésped haga cambios o tenga peticiones especiales indica que probablemente no cancele.




previous cancellations

El 78% de los huéspedes que repiten, no cancelan.

Fidelizar ¿?




Variables Feature Importance



transient
transient-party

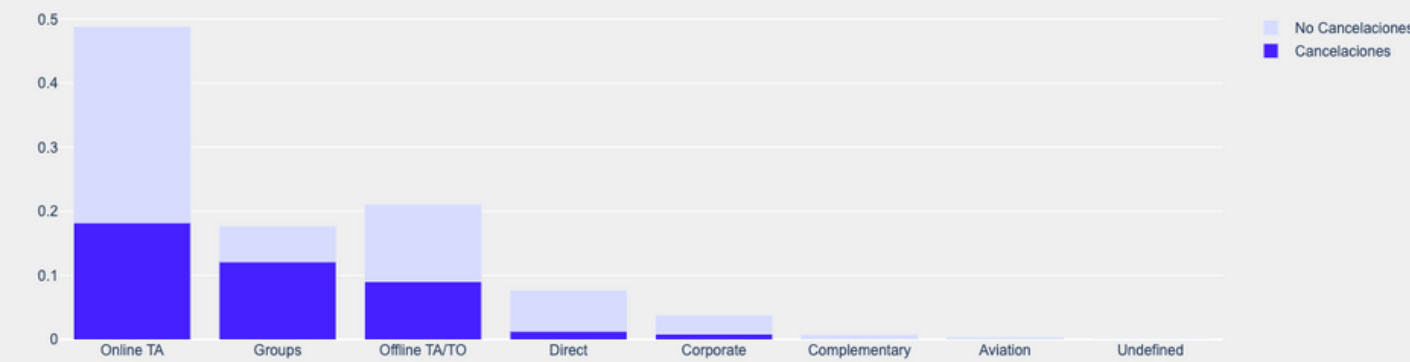
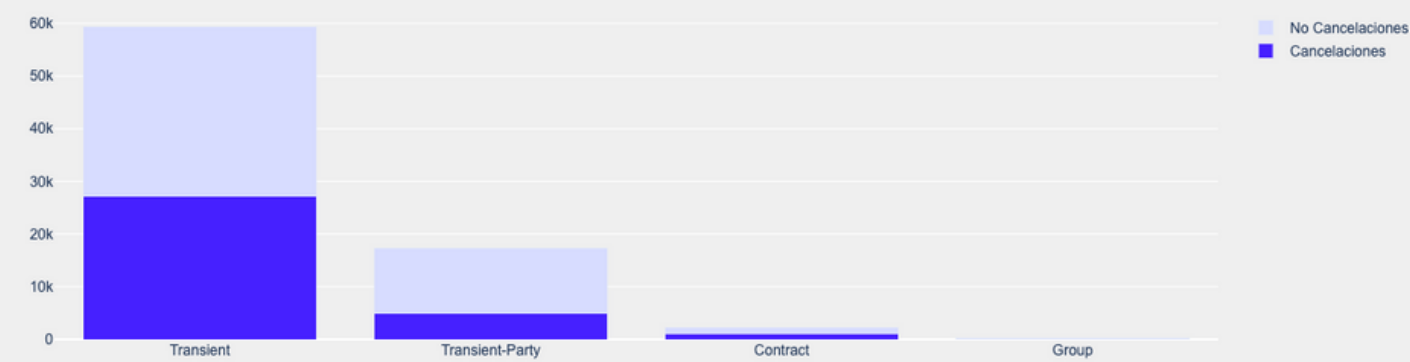
Reservas Individuales

Casi todas las reservas son de este tipo, no nos da mucha información esta columna.



market segment

Destaca la categoría 'Groups', con más cancelaciones que reservas sin cancelar.



Variables Feature Importance

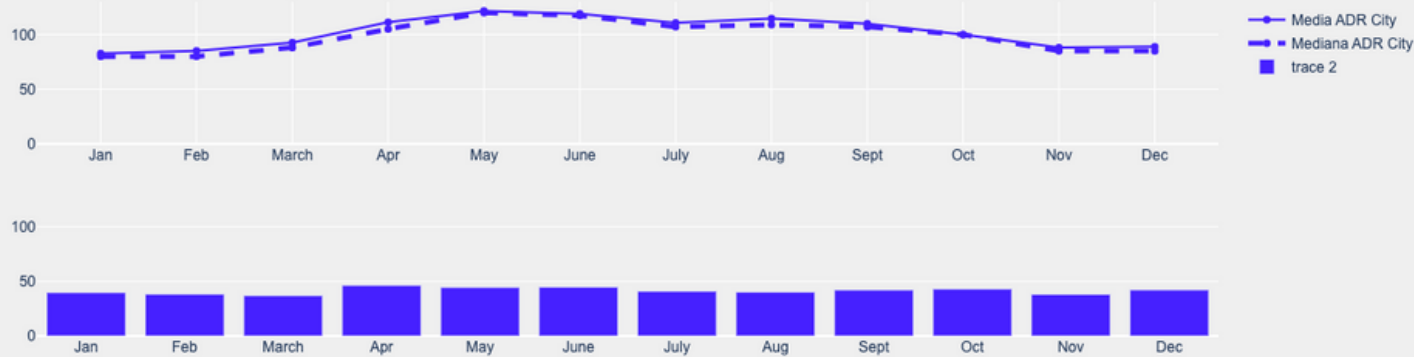


arrival date month

arrival date day of month

El día y el mes de la fecha de reserva.

Destacamos que al ser un hotel de ciudad, no hay picos de precio tan destacados como en los hoteles con estacionalidad.



ML

GRACIAS

SUPERVISADO

1
0