# Reservaciones de hotel

Análisis de datos con Phyton

Perla Conchita Peña Campos

# Industria hotelera Portugal

## 4, 983 millones de euros

VENTAS DE ALOJAMIENTO 2018

## 67%

PROVIENE DE LA INDUSTRIA HOTELERA

# Gestión hotelera

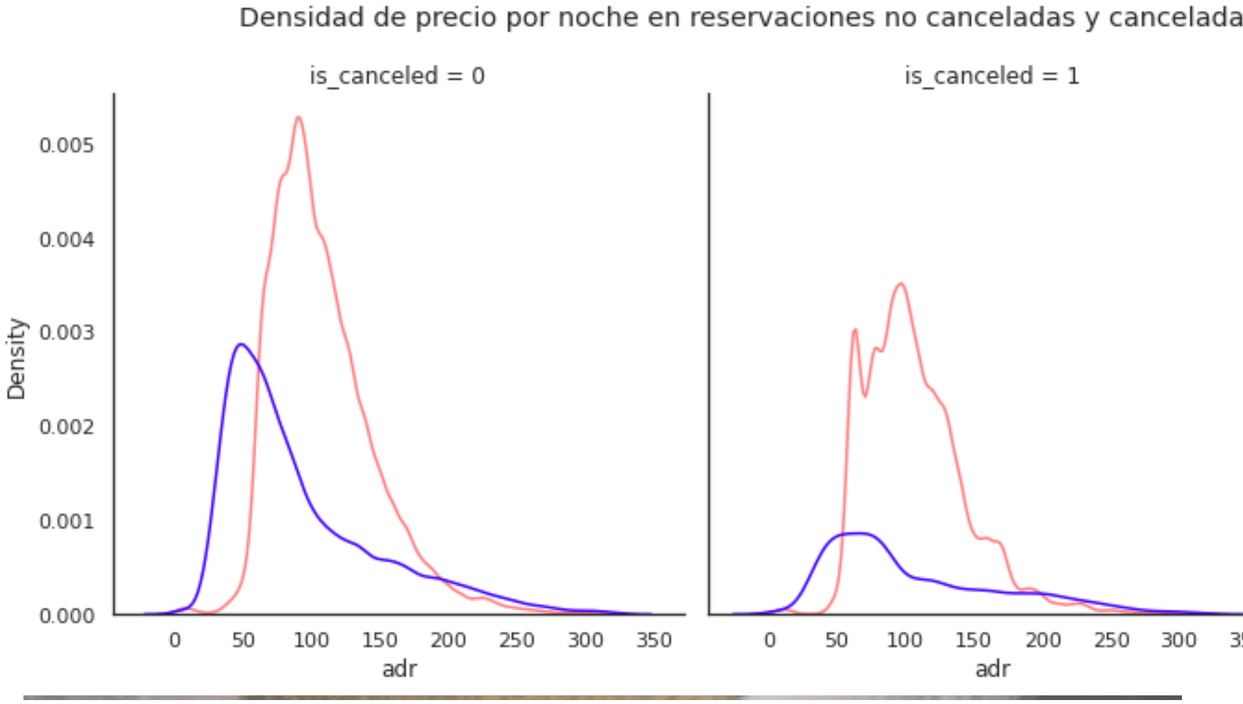
PRINCIPAL PROBLEMA

Fuente: https://www.icex.es/icex/es/navegacion-principal/todos-nuestros-servicios/informacion-de-mercados/sectores/servicios/documentos/DOC2019825728.html

# Contenido

El data set contiene información de dos hoteles de Portugal, resort (H1) y de ciudad (H2). Cuenta con 31 variables que describen 40,060 observaciones para el de resort y 79,330 observaciones para el de ciudad (H2). Comprende información del 01 julio del 2015 al 31 de agosto del 2017

.https://www.sciencedirect.com/science/article/pii/S2352340918315191




## Estimados de locación y variabilidad

Analizar las variables y eliminar valores atípicos (dos métodos).

## EDA

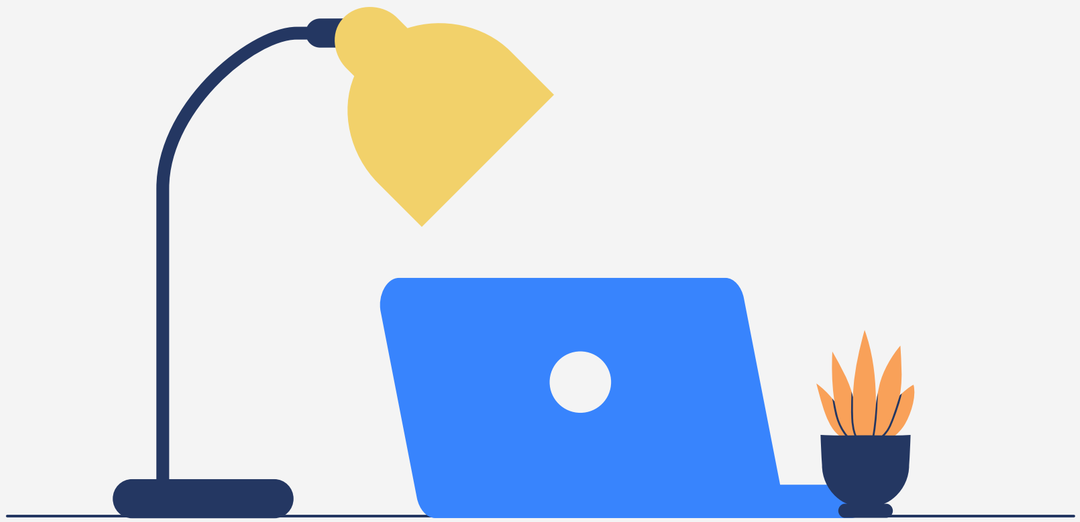
1. Comportamiento hotelero (4)
2. Sobre los hoteles (6)
3. Canales de distribución (3)
4. Análisis cancelaciones (8)

## Correlación de variables

1. Entre las variables del data set
2. Correlación con "is_canceled"

# Modelo predictivo

Predecir la cancelación de una reservación



## Training test split y preprocessing

1. Data set de test y train
2. Preprocessing test y train (Pipeline):
    a. StandarScalar (numéricas)
    b. OneHotEncoder (categóricas)

## Modelos, curva ROC/AUC, validación

1. Regresión logística
2. SVM
3. Kmeans (para práctica)

## Conclusiones

1. Modelo
2. Personales

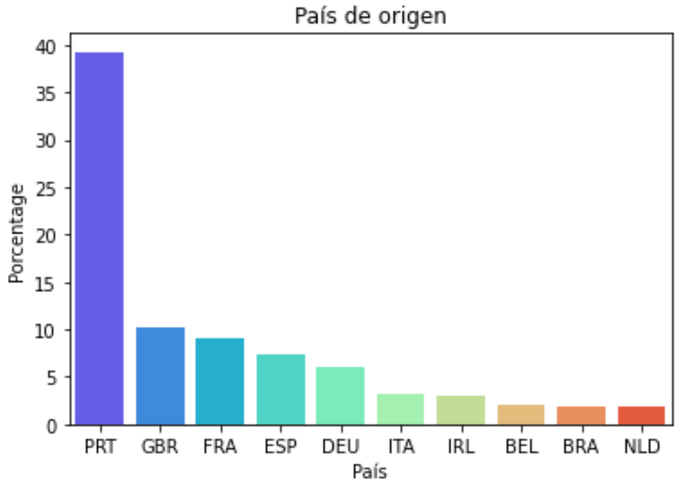
# EDA

## Resumen

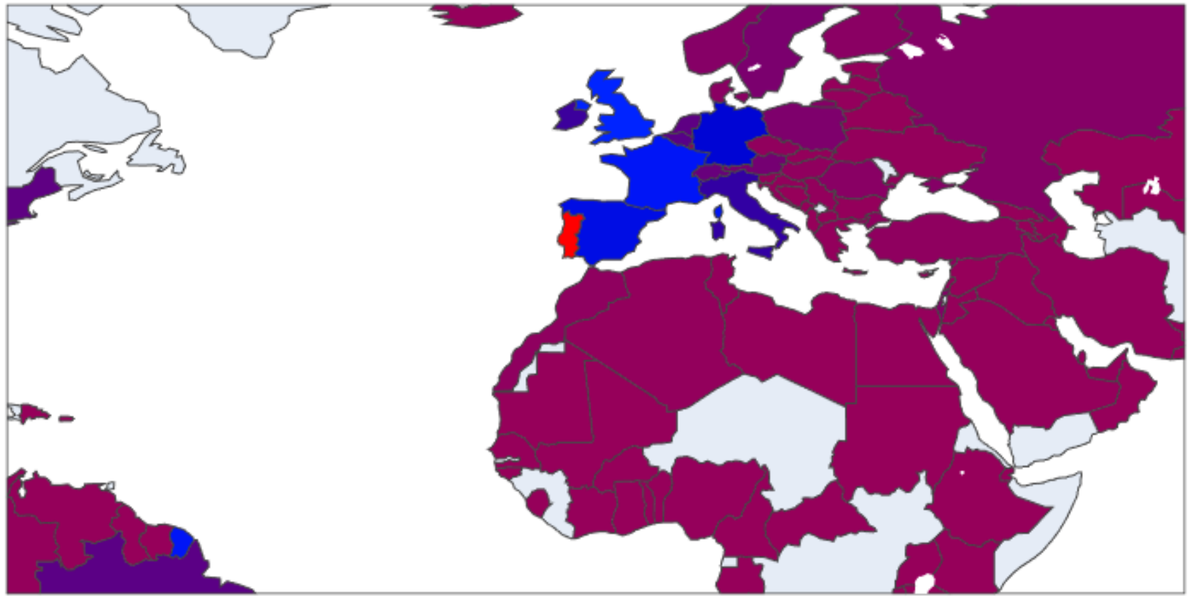
# EDA

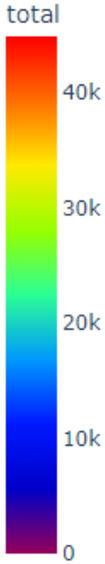
## Resumen

# País de procedencia



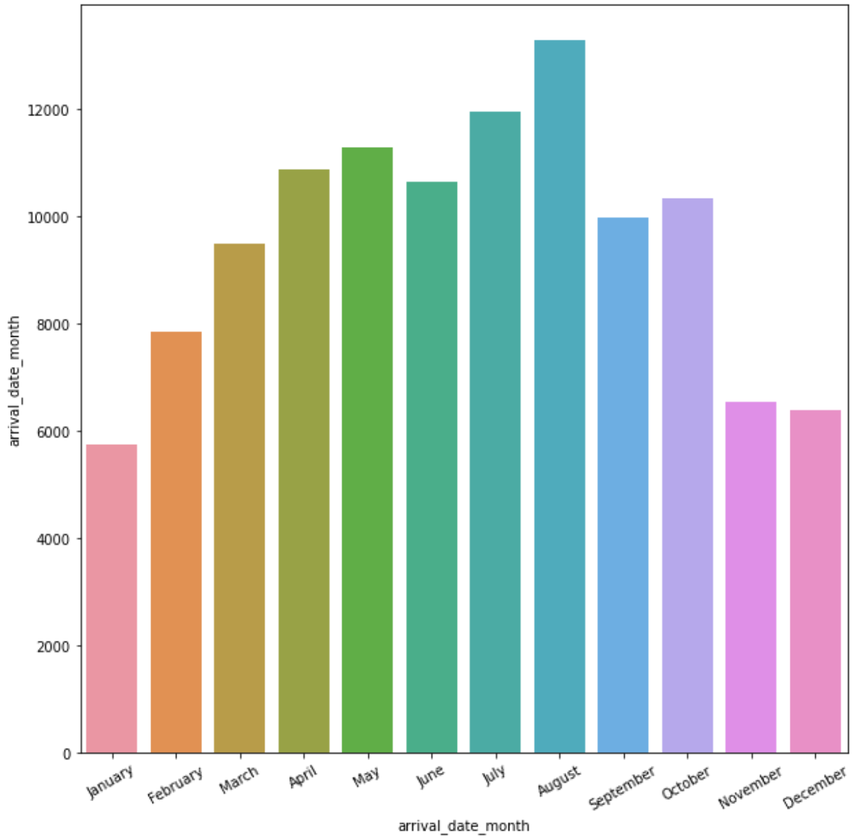

Cantidad de personas



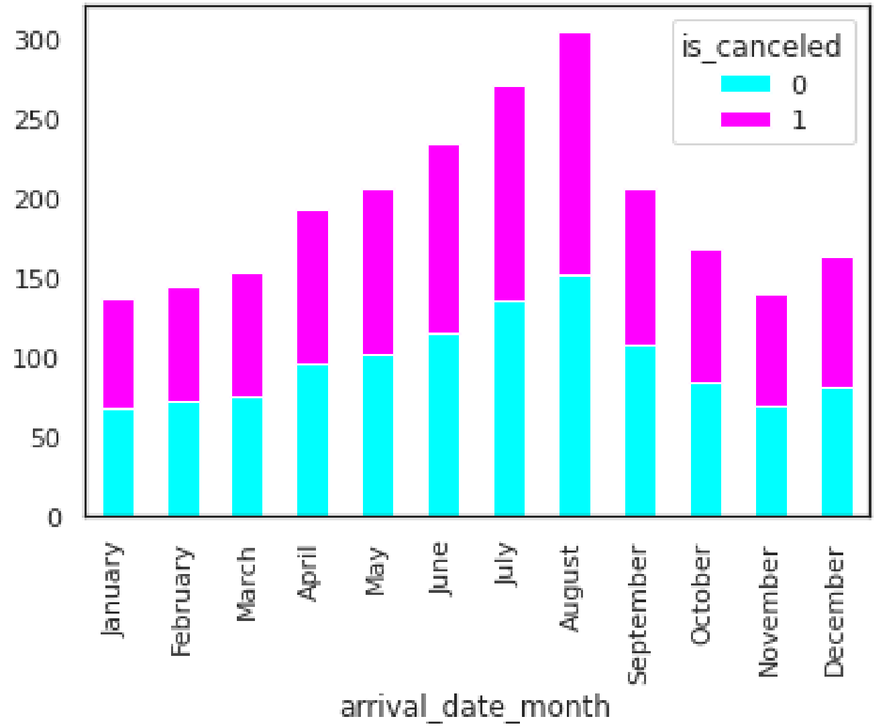


# Mes de llegada

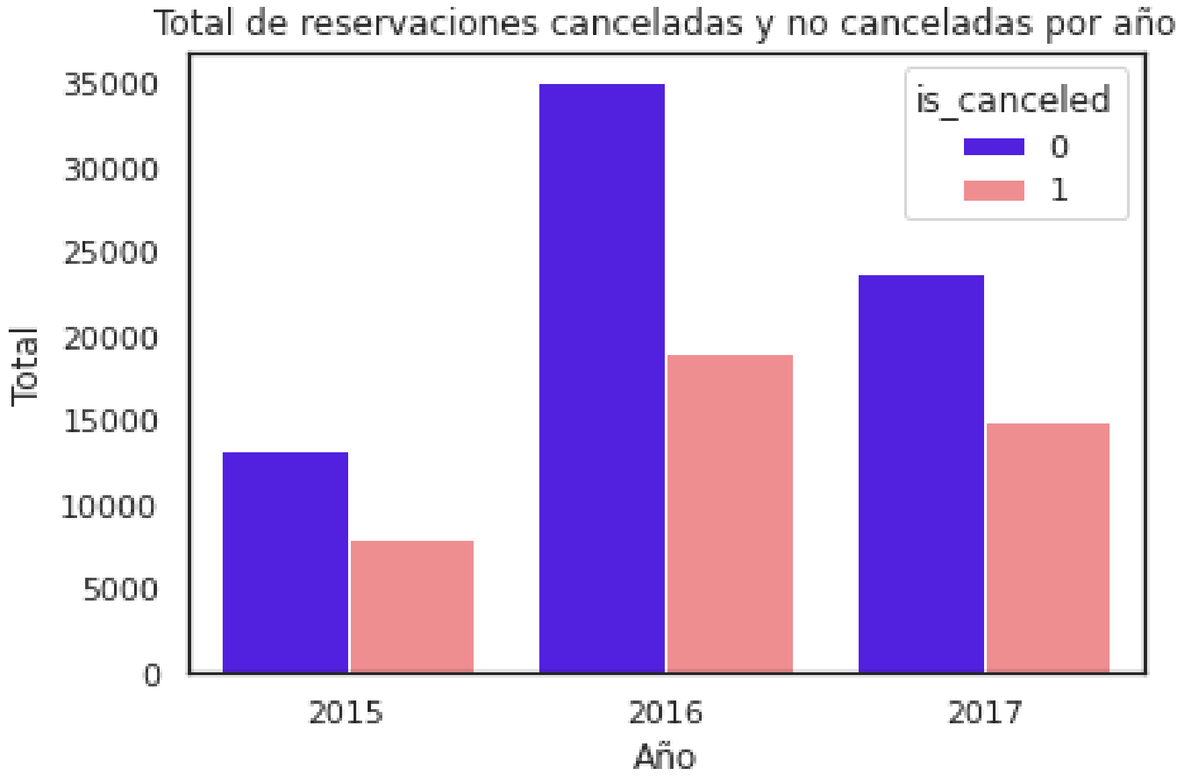
Cantidad de reservaciones por mes



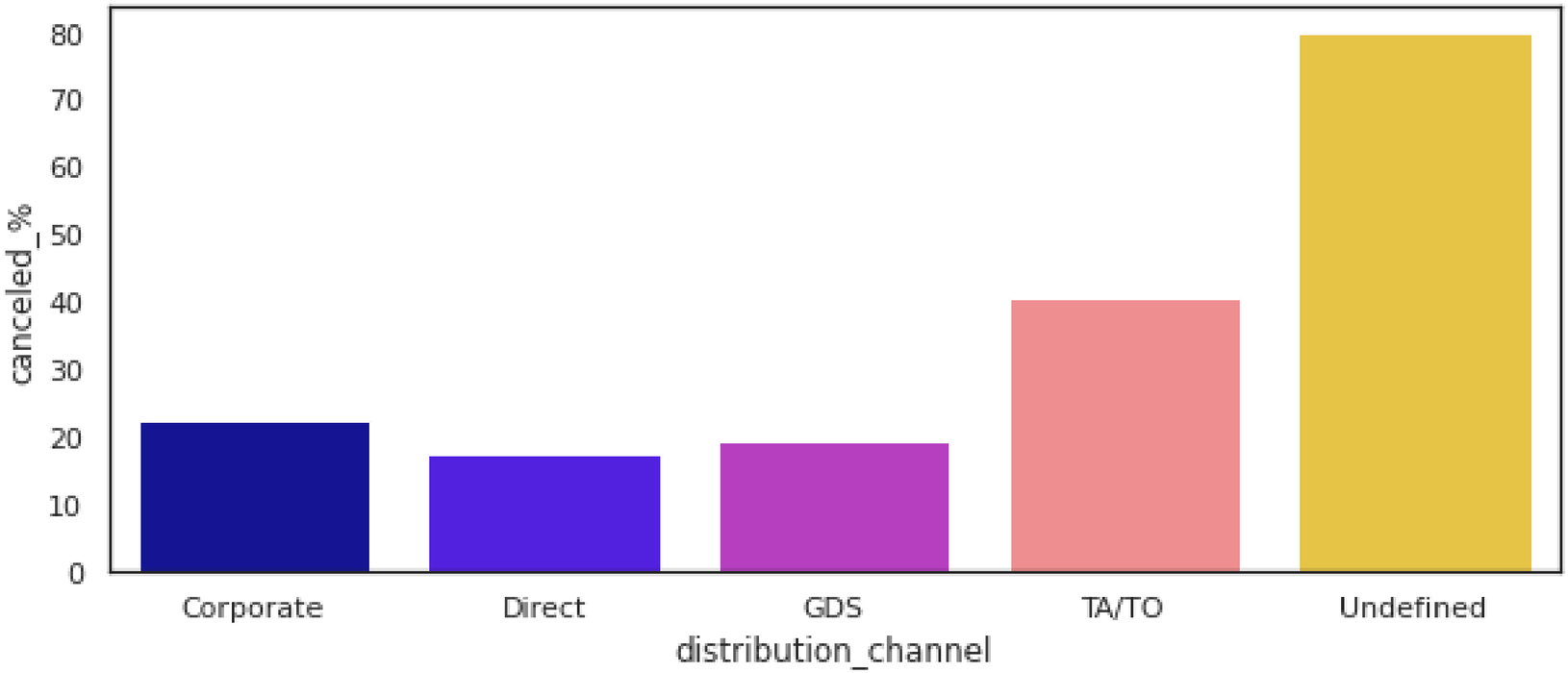

Precio promedio por noche mensual



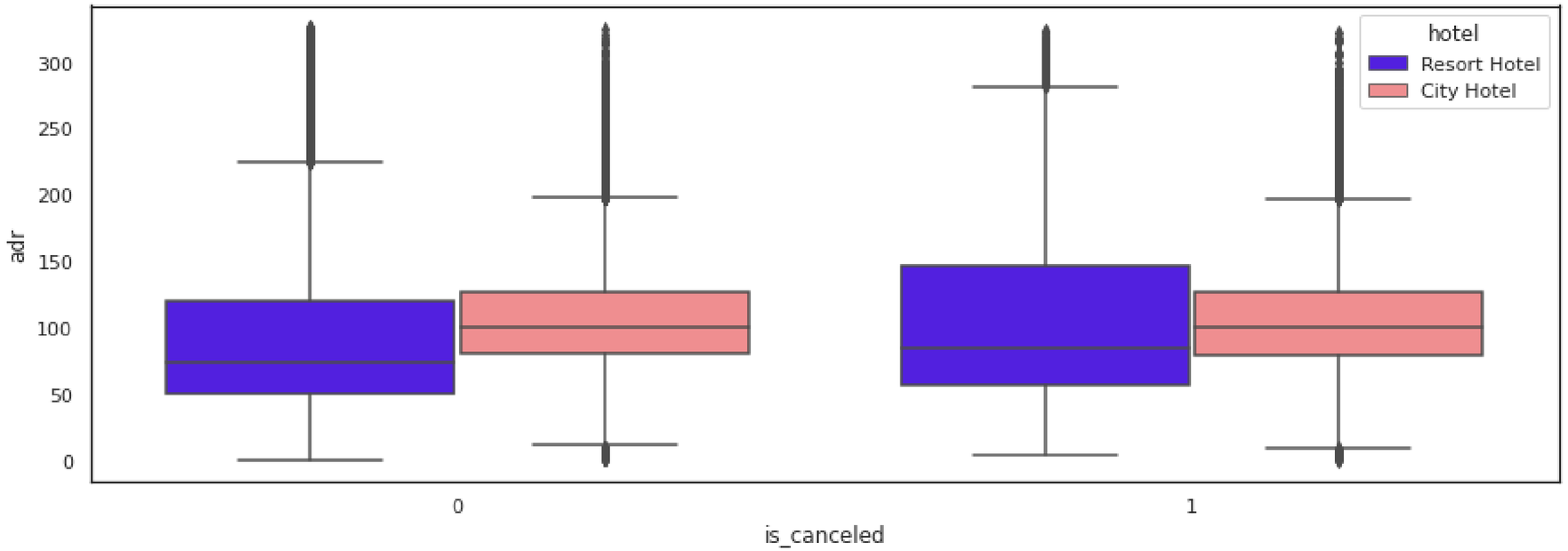
# Cancelacion es por año



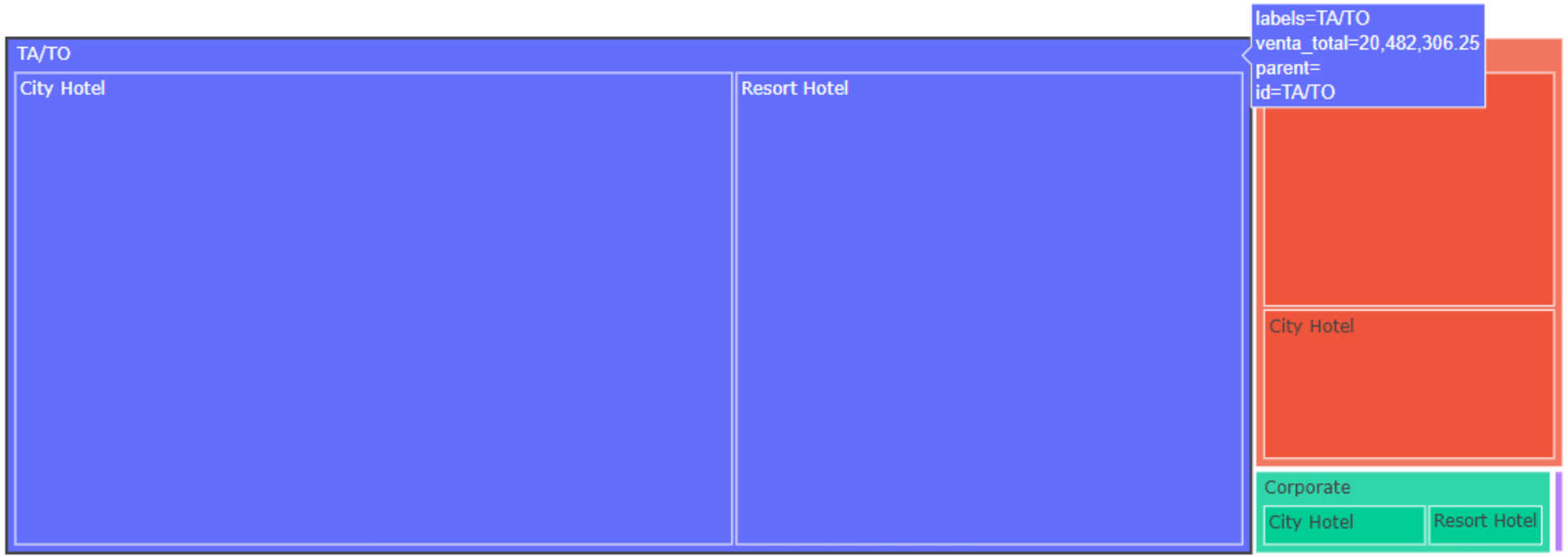
# Cancelacion es por canal



# Precio promedio de reservaciones



# Canal con mayores ventas en reservaciones efectivas



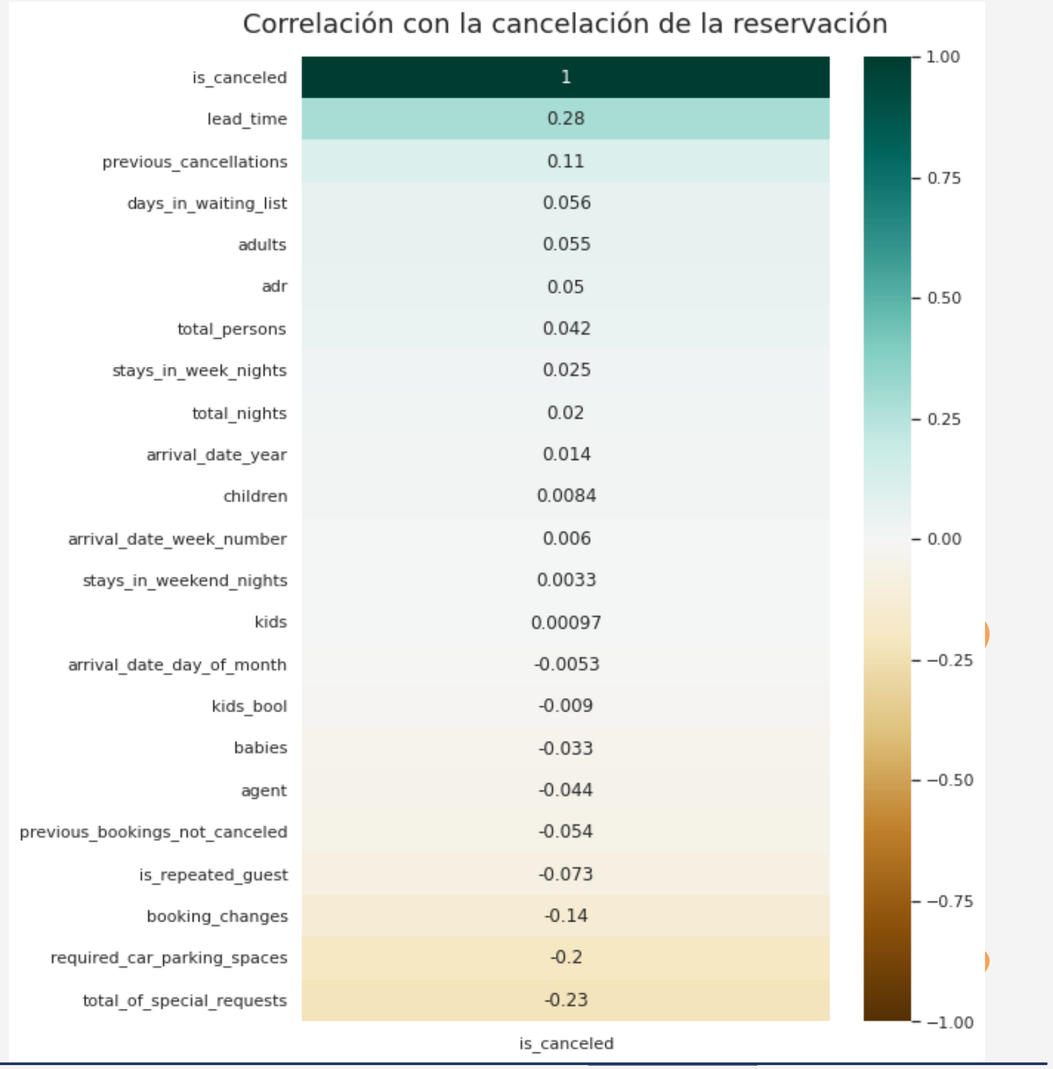
# Correlación

# Variables numéricas

```
['lead_time',
 'stays_in_weekend_nights',
 'stays_in_week_nights',
 'adults',
 'children',
 'babies',
 'is_repeated_guest',
 'previous_cancellations',
 'previous_bookings_not_canceled',
 'adr',
 'required_car_parking_spaces',
 'total_of_special_requests']
```

# Variables categóricas

```
['hotel',
 'arrival_date_month',
 'meal',
 'market_segment',
 'distribution_channel',
 'reserved_room_type',
 'deposit_type',
 'customer_type']
```

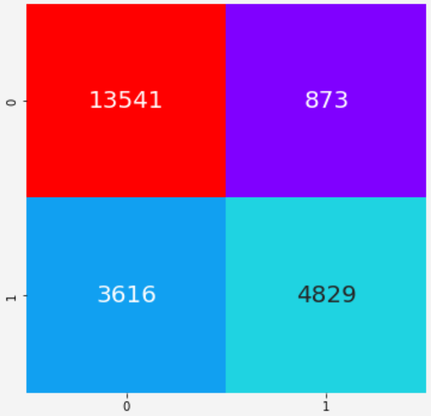


# Modelos

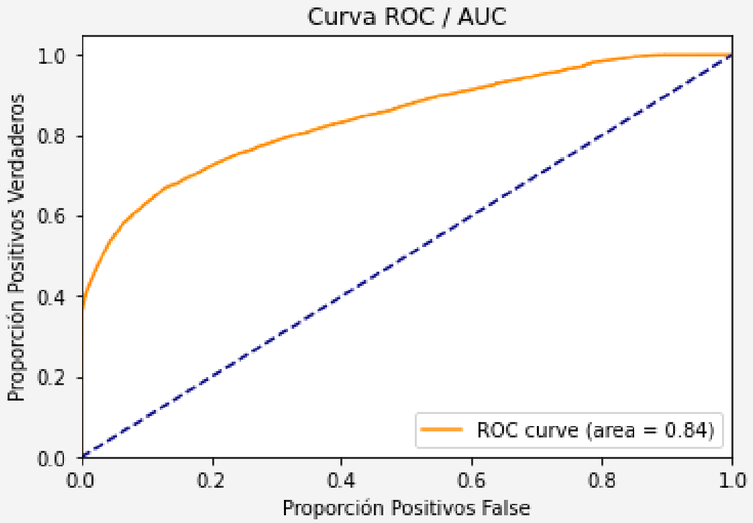
Predicción de las reservaciones

# Regresión logística

Matriz de confusión




```
Precision: 0.8468958260259558
Exactitud: 0.8036222056957872
Sensibilidad: 0.5718176435760806
Especificidad: 0.939433883724157
```




## Python

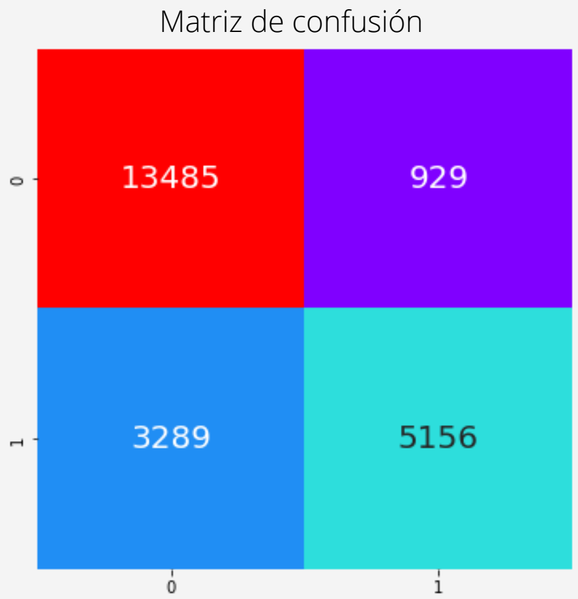
1. Precisión: que tan cerca esta el resultado de una predicción del valor verdadero.
2. Exactitud: Porcentaje de predicciones correctas del total realizadas.
3. Sensibilidad: Es la tasa de verdaderos positivos, es decir, la proporción de casos positivos bien clasificados por el modelo respecto al total.
4. Especificidad: Tasa de verdaderos negativos; la proporción de negativos bien clasificados por el modelo.

```
                                Pr(>|z|)
(Intercept)                      < 2e-16 ***
deposit_typeNon Refund           < 2e-16 ***
deposit_typeRefundable          0.198390
lead_time                        < 2e-16 ***
market_segmentComplementary     0.037156 *
market_segmentCorporate         0.000517 ***
market_segmentDirect            6.59e-06 ***
market_segmentGroups            0.130061
market_segmentOffline TA/TO     4.41e-06 ***
market_segmentOnline TA         0.081027 .
market_segmentUndefined         0.829173
adr                              < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

## RStudio

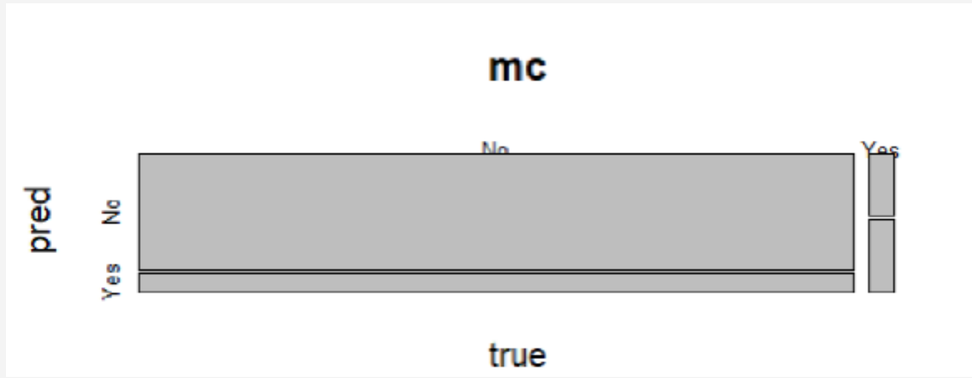
Identificar las variables con mayor peso para el modelo

# SVC




```
Precision: 0.847329498767461
Exactitud: 0.8154774924537381
Sensibilidad: 0.6105387803433985
Especificidad: 0.939433883724157
```

## Phyton



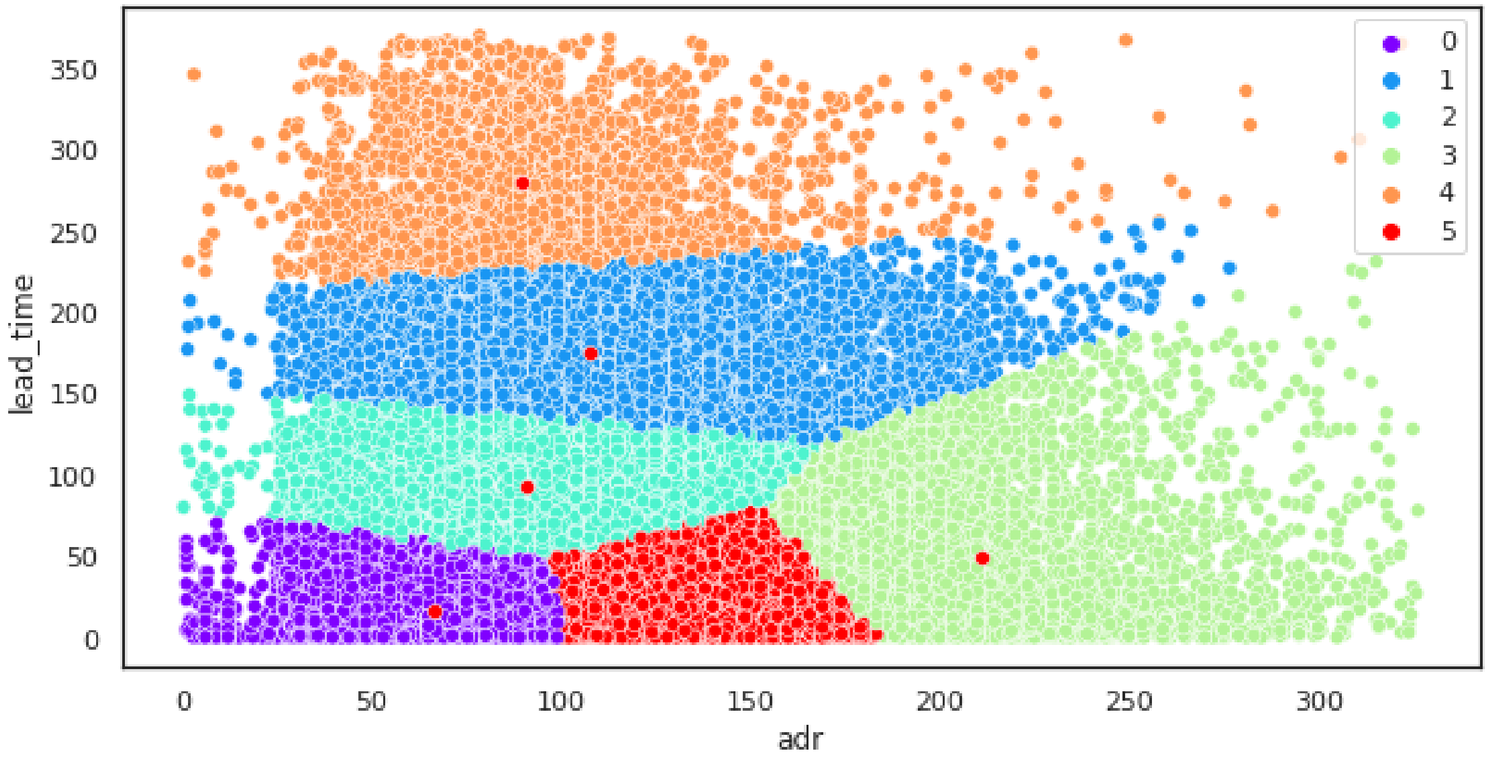

```
>mc
pred
true    No  Yes
  No  4163  657
  Yes   82   98

#accurancy
> round(sum(diag(mc))/sum(colSums(mc)), 5)
[1] 0.8522
```

## RStudio

Se puede optimizar utilizando GridSearchCV

# KMeans

Clusters de acuerdo al precio promedio por noche y tiempo de espera



# Conclusiones

Repositorio