UNIVERSIDAD
Panamericana

Materia: Inteligencia de Negocio y Soluciones de Ciencia de Datos

Profesor: Edgar Avalos Gauna

Alumnos:

Jonatan Chirinos Funatsu

Daniel Hernández Toledo

Miranda Galván Medina

Teresa Sofía González Morales

Sebastián Cruz Hernández

Trabajo: Proyecto Final

Fecha de entrega: 13/12/2025

Índice

Índice **2**
Propuesta de Negocio: Optimización de ingresos y cancelaciones en la Industria Hotelera **3**
Objetivos **5**
1. Optimización de precios y detección de oportunidades de ingresos 5
2. Reducción de cancelaciones 6
EDA (Análisis exploratorio) **6**
- Identificar tipos de datos 6
- Identificar valores atípicos 6
- Identificar valores faltantes 6
- Identificar métricas estadísticas 6
Gráficas: 8
Modelo de entrenamiento **21**
Resultados y conclusiones **28**
Plan de Acción a Implementar **30**

Propuesta de Negocio: Optimización de ingresos y cancelaciones en la Industria Hotelera

Para este proyecto proponemos utilizar un dataset de reservas hoteleras en Portugal para desarrollar un modelo predictivo que permita mejorar las estrategias de gestión de ingresos tanto en hoteles urbanos como en hoteles tipo resort, así también se realizó un modelo para analizar las cancelaciones y poder indagar más en las razones por las que los clientes cancelaron sus reservaciones. A través del análisis detallado del comportamiento de tarifas, estacionalidad, patrones de reserva y características de los clientes, donde nuestro objetivo principal es encontrar oportunidades para incrementar ingresos y reducir la tasa de cancelaciones en la industria hotelera.

El dataset empleado fue extraído de la siguiente fuente:

(https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand).

Features del dataset:

(siguiente hoja)

Feature	Descripción
Hotel	Nos dice el tipo de hotel “Resort Hotel” o “City Hotel”
is_canceled	Nos indica si la reservación fue cancelada
lead_time	Número de días entre la fecha de reserva y la fecha de llegada
arrival_date_year	Año de la fecha de llegada
arrival_date_month	Mes de la fecha de llegada
arrival_date_week_number	Número de semana del año para la fecha de llegada
arrival_date_day_of_month	Día de la fecha de llegada
stays_in_weekend_nights	Número de noches de fin de semana (sábado o domingo) que el huésped se alojó o reservó para alojarse en el hotel
stays_in_week_nights	Número de noches de semana (de lunes a viernes) que el huésped se alojó o reservó para alojarse en el hotel
adults	Número de adultos en la reserva
children	Número de niños en la reserva
babies	Número de bebés en la reserva
meal	Tipo de comida reservada. Las categorías se presentan en paquetes de comidas estándar de hospitalidad: Sin definir/SC: sin paquete de comidas; AD: alojamiento y desayuno; MP: media pensión (desayuno y otra comida, generalmente cena); PC: pensión completa (desayuno, almuerzo y cena).
country	País de origen
market_segment	Designación del segmento de mercado.
distribution_channel	Canal de distribución de reservas.
is_repeated_guest	Valor que indica si el nombre de la reserva es de un huésped repetido (1) o no (0)
previous_cancelations	Número de reservas anteriores que fueron canceladas por el cliente antes de la reserva actual
previous_bookings_not_canceled	Número de reservas anteriores no canceladas por el cliente antes de la reserva actual
reserved_room_type	Código del tipo de habitación reservada.
assigned_room_type	Código del tipo de habitación asignada a la reserva.
booking_changes	Número de cambios/modificaciones realizadas en la reserva desde el momento en que se introdujo la reserva en el PMS hasta el momento del check-in o cancelación
deposit_type	Indica si el cliente realizó algún depósito para garantizar la reserva. Puede ser de tres maneras, Sin depósito, Sin reembolso, Reembolsable
agent	ID de la agencia de viajes que realizó la reserva
company	ID de la empresa/entidad que realizó la reserva o responsable del pago de la reserva.
days_in_waiting_list	Número de días que la reserva estuvo en lista de espera antes de ser

	confirmada al cliente
customer_type	Tipo de reserva, asumiendo una de cuatro categorías: Contrato, Grupo, Transitoria o Parte transitoria.
adr	Tarifa diaria promedio, definida como la división de la suma de todas las transacciones de alojamiento por el número total de noches de estadía.
required_car_parking_spaces	Número de plazas de aparcamiento solicitadas por el cliente
total_of_special_request	Número de peticiones especiales realizadas por el cliente
reservation_status	Último estado de la reserva, asumiendo una de tres categorías: Cancelada: el cliente canceló la reserva; Salida: el cliente se registró pero ya se fue; No presentación: el cliente no se registró y no informó al hotel el motivo.
reservation_satatus_date	Fecha en la que se estableció el último estado.

Analizando este conjunto de datos tendremos un mejor conocimiento del desempeño tarifario, analizando la variable de ADR (Average Daily Rate), que refleja el ingreso promedio por habitación ocupada igualmente el dataset resulta clave para el análisis del comportamiento de cancelaciones, variable fundamental para evaluar el impacto de las reservas en los ingresos del hotel.

Si bien el dataset original cuenta con un número elevado de *features*, no todas fueron utilizadas para el análisis ni para el entrenamiento de los modelos ya que algunas variables no tienen ninguna correlación con las variables objetivos del estudio desempeño tarifario y predicción de cancelaciones, por lo cual no aportan valor predictivo, además en el caso de las variables de fecha, se aplicó una reducción de dimensionalidad, como se detalla en el apartado de EDA precisamente para este objetivo.

Objetivos

Los objetivos planteados para este proyecto son el de optimización de precios y la reducción de cancelaciones ya que ambos consideramos que son complementarios cuando hablamos de ingresos para los hoteles, puesto que con el ADR mediremos la capacidad del hotel para optimizar sus ingresos y el análisis de las cancelaciones nos ayudará a evaluar la estabilidad de los ingresos, ya que pues una tarifa pierde su valor si las reservas no se completan. A continuación se explica a mayor detalle cada objetivo.

1. Optimización de precios y detección de oportunidades de ingresos

El análisis temporal del ADR por mes, tipo de hotel y segmento de cliente permitirá identificar:

- Meses con alta disposición a pagar (picos tarifarios naturales).
- Meses donde el ADR se encuentra por debajo del potencial del mercado.
- Diferencias entre hoteles tipo “City” y “Resort” para ajustar estrategias diferenciadas.
- Impacto de variables como país de origen, duración de la estancia o distribución del mercado.

2. Reducción de cancelaciones

El análisis del riesgo de cancelación permitirá identificar:

- Qué segmentos tienen mayor probabilidad de cancelar.
- Cómo influyen las políticas de pago y depósito en las cancelaciones.
- Qué factores (anticipación de la reserva, tipo de habitación, canal) predicen cancelaciones.
- Estrategias para minimizar el impacto en ocupación e ingresos.

EDA (Análisis exploratorio)

Cabe aclarar que se realizaron dos análisis exploratorios independientes, uno para cada objetivo del proyecto: optimización de precios (ADR) y predicción de cancelaciones. Aunque la limpieza básica de datos es similar en ambos casos (eliminación de nulos, validación de valores ilógicos), el análisis de relaciones entre variables, la selección de features y las visualizaciones son específicas para cada problema. Esto se debe a que cada modelo tiene necesidades diferentes y las mismas features pueden aportar valor predictivo distinto dependiendo de la variable objetivo, por lo que la selección final de características no es idéntica entre ambos notebooks.

Para el EDA, se realizaron los siguientes puntos:

- Identificar tipos de datos

Observar el tipo de variable que tienen asignados los features para poder preparar mejor el proceso pre entrenamiento del modelo y así poder tener un buen resultado. En el caso de features con un tipo de variable incorrecto o ambiguo se realiza el cambio al tipo de variable correcto.

- Identificar valores atípicos

Revisar valores atípicos (o outliers) ayuda a detectar instancias que salen del comportamiento normal en comparación a los otros datos. Estos registros pueden afectar el entrenamiento del modelo agregando ruido e insertando valores que no representan un valor real de nuestros datos. En este apartado se explica lo que se decidió hacer para cada valor atípico encontrado.

- Identificar valores faltantes

Detectar valores faltantes (como null o NaN) permite saber qué tan completo está el dataset. Según la cantidad y relevancia del feature podemos decidir cómo manipular este tipo de instancias. Este tratamiento evita errores posteriores al trabajar tanto con gráficas como con el entrenamiento del modelo

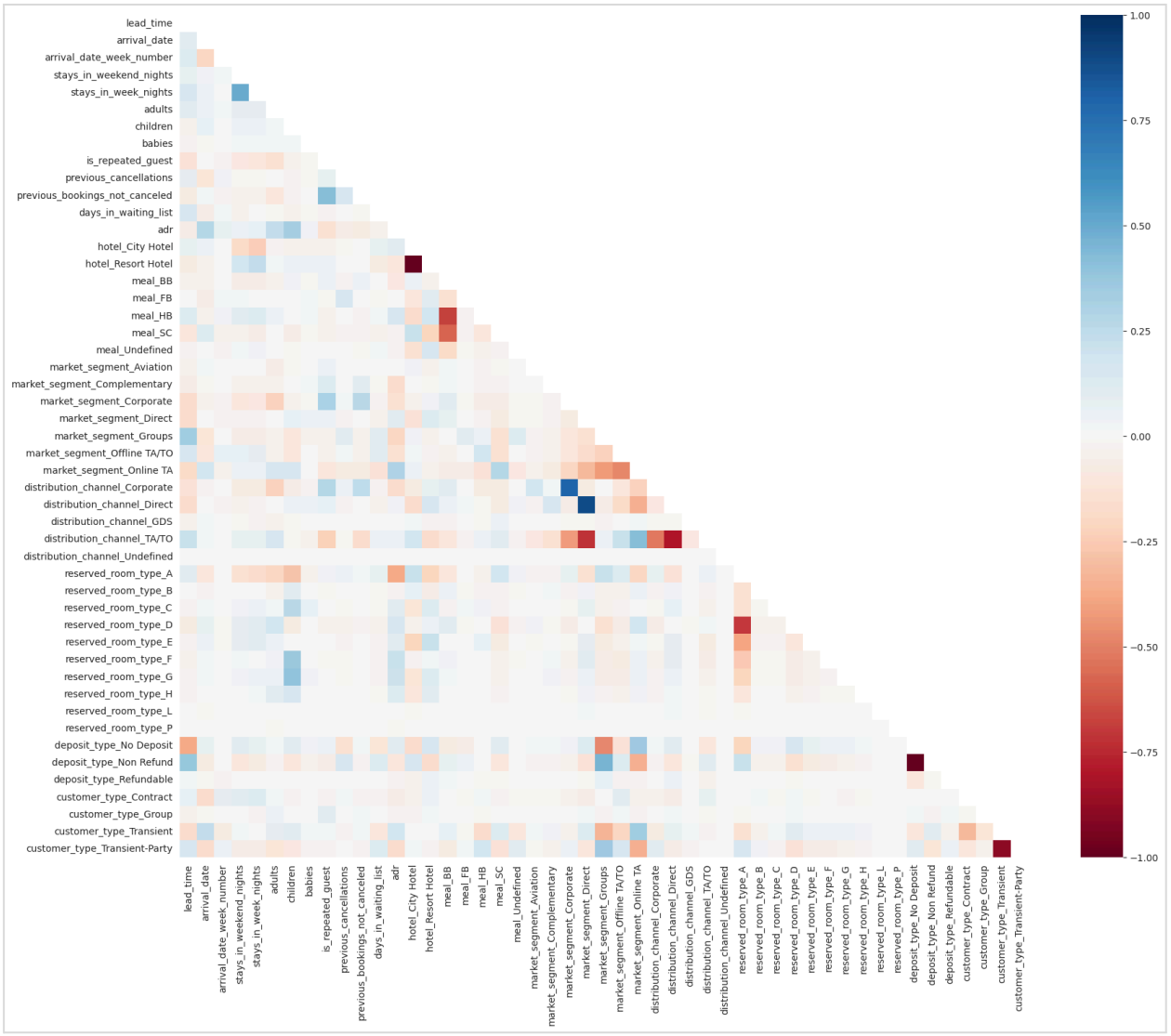
- Identificar métricas estadísticas

Calcular métricas básicas como media, mediana, desviación estándar o percentiles ayuda a entender mejor la distribución de los datos. Esto facilita identificar patrones, posibles problemas y confirmar si las variables están listas para el modelado. También sirve para tomar decisiones más informadas durante el preprocesamiento y hasta con la identificación de valores atípicos que en ocasiones no son fáciles de detectar.

Así también, es necesario hacer un análisis visual de los datos para entender mejor su distribución, e identificar patrones y tendencias que nos den una idea más clara del comportamiento del dataset. Esto incluye observar cómo se relacionan entre sí los diferentes features, apoyándonos en distintos tipos de gráficas para obtener una visión más completa.

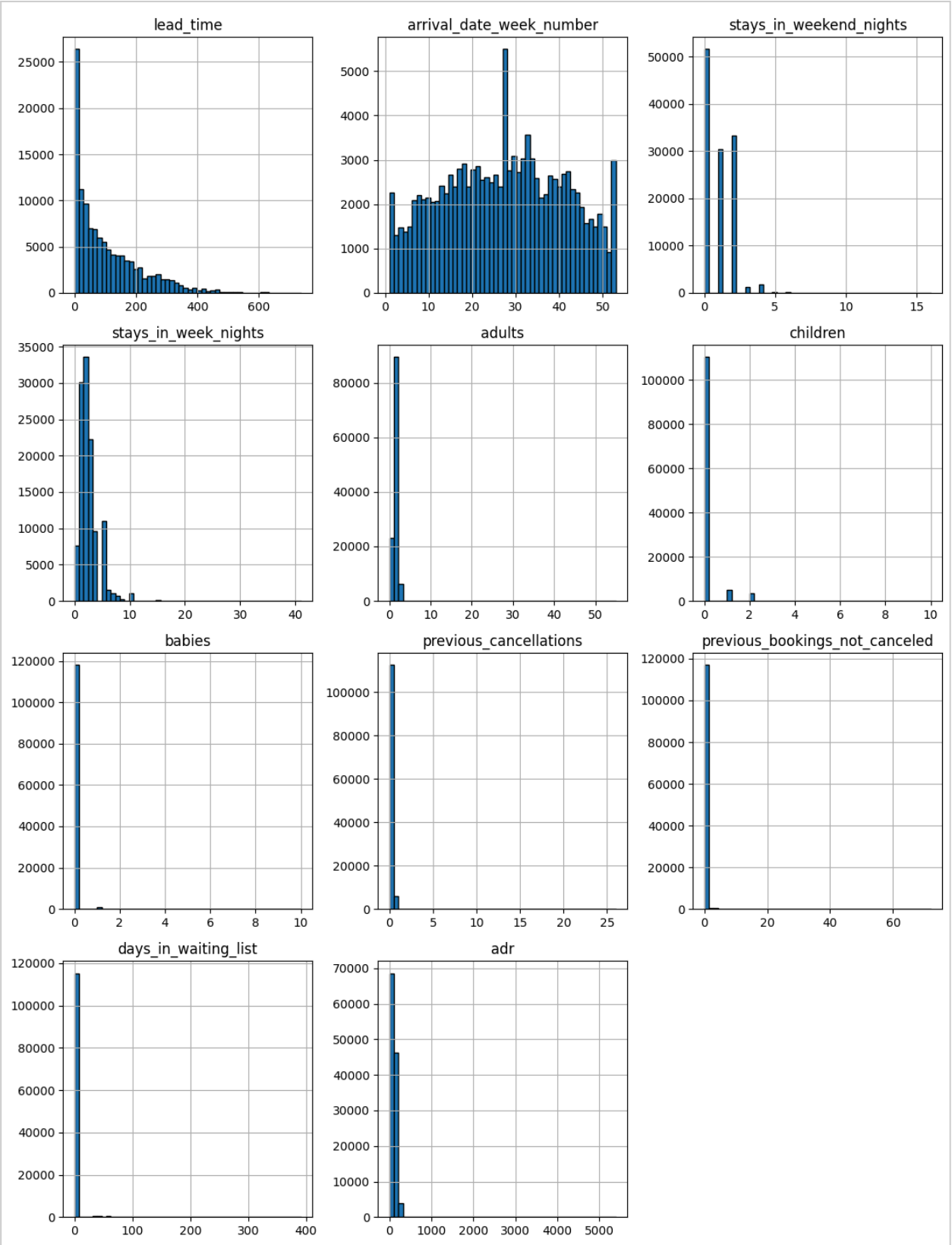
Gráficas:

Matriz de correlación



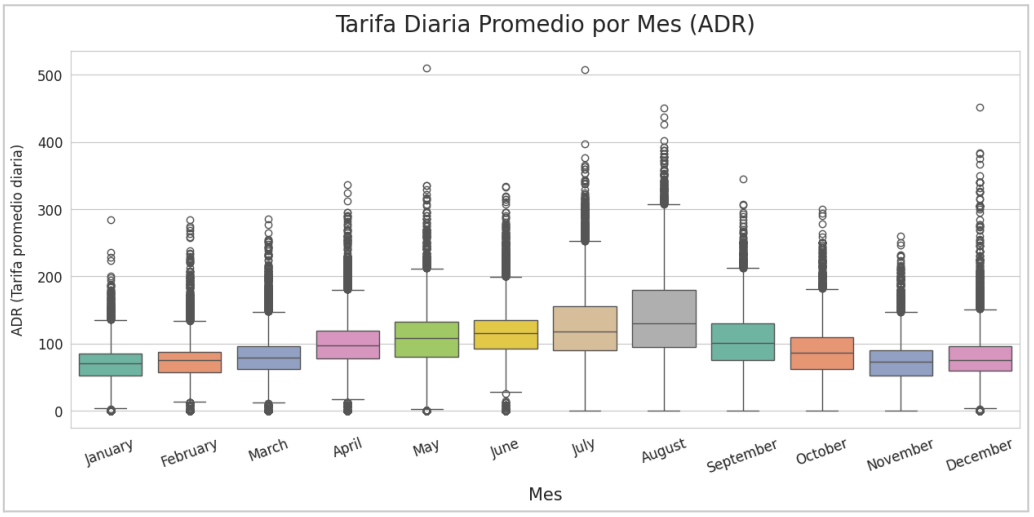
En primer lugar utilizamos una matriz de correlación para poder ver la manera en que se relacionan los diferentes features, esto es crucial para poder detectar los patrones y redundancias dentro del dataset.

Histograma de variables numéricas

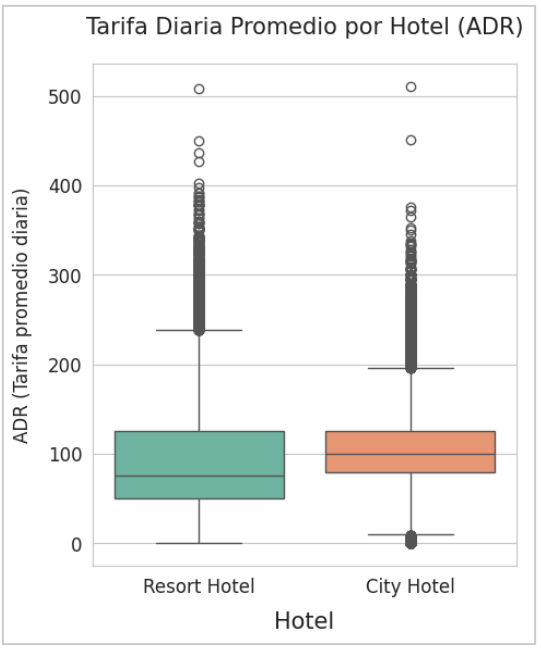


Para las variables numéricas, utilizamos histogramas con el fin de examinar la distribución y dispersión de los datos. Esta visualización nos permitió identificar valores atípicos que podrían requerir una limpieza antes del modelado

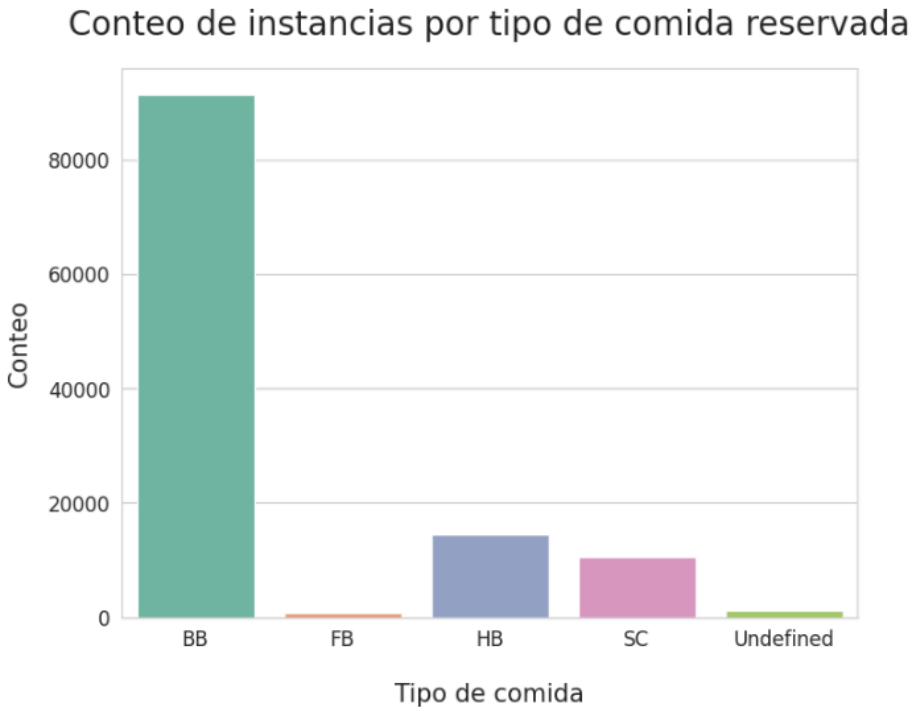
Objetivo 1: Optimización de precios y detección de oportunidades de ingresos



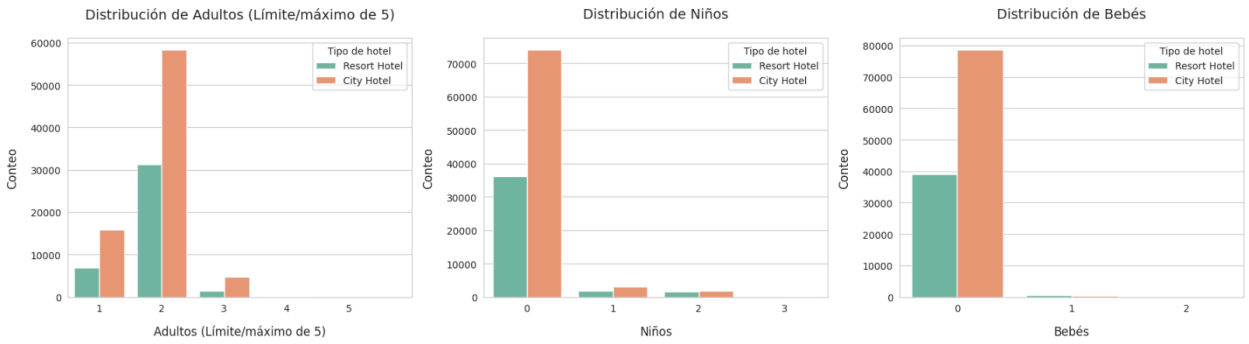
Esta gráfica nos permite hacer algunas observaciones como que la tarifa promedio está entre los 50 y 125, teniendo una alza principalmente entre los meses de junio, julio y agosto, es decir, un alza en la demanda en los meses de verano.



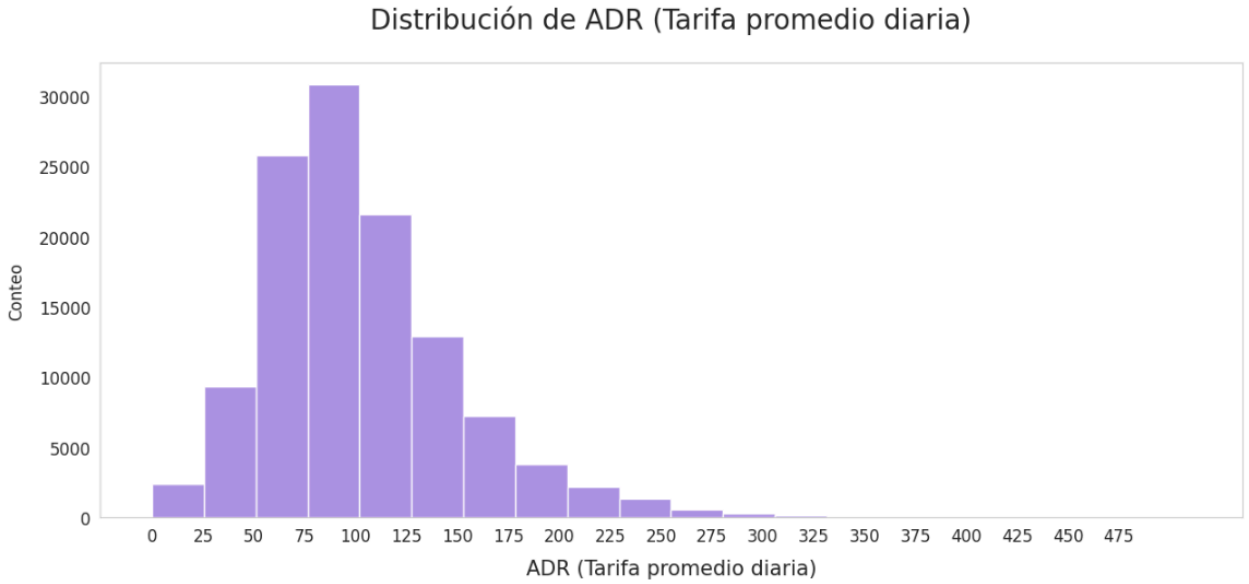
Esta gráfica nos deja ver como varían las tarifas para ambos tipos de hotel. Así, los hoteles de ciudad presentan tarifas promedio más altas y más estables, mientras que los resorts muestran una mayor variabilidad en precios, con tarifas que fluctúan significativamente dependiendo del contexto.



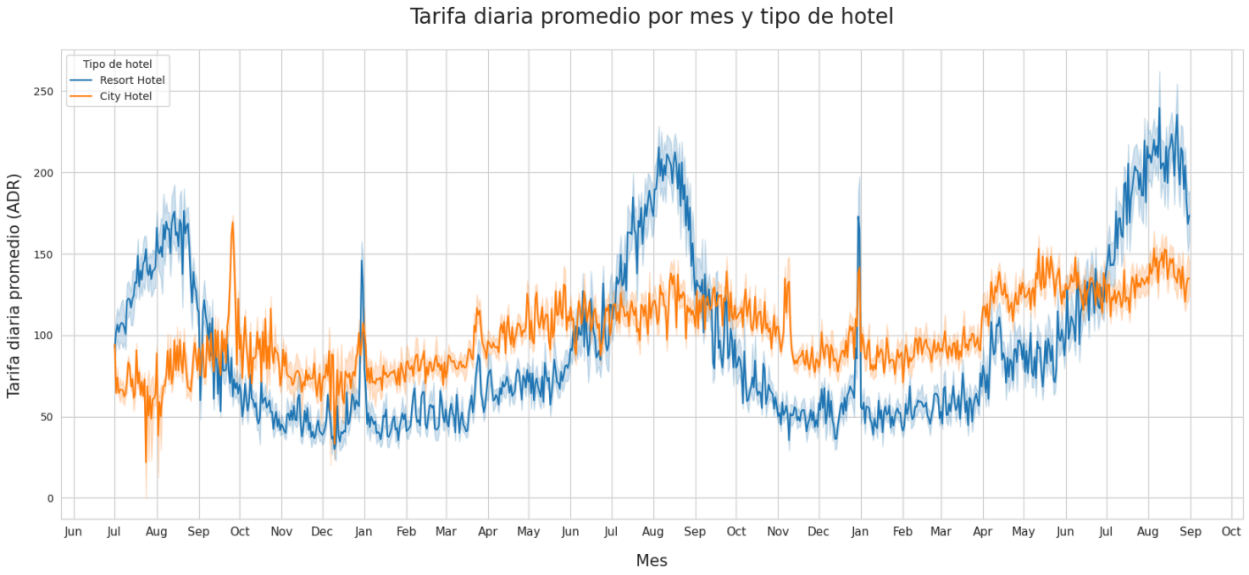
Entre el tipo de comida reservada, podemos observar que la opción más popular es la de bed & breakfast, seguido de half board y self-catering. Es interesante ver cómo los clientes prefieren esta opción o de alguna manera como los hoteles priorizan esta opción de comida a la hora de hacer reservaciones.

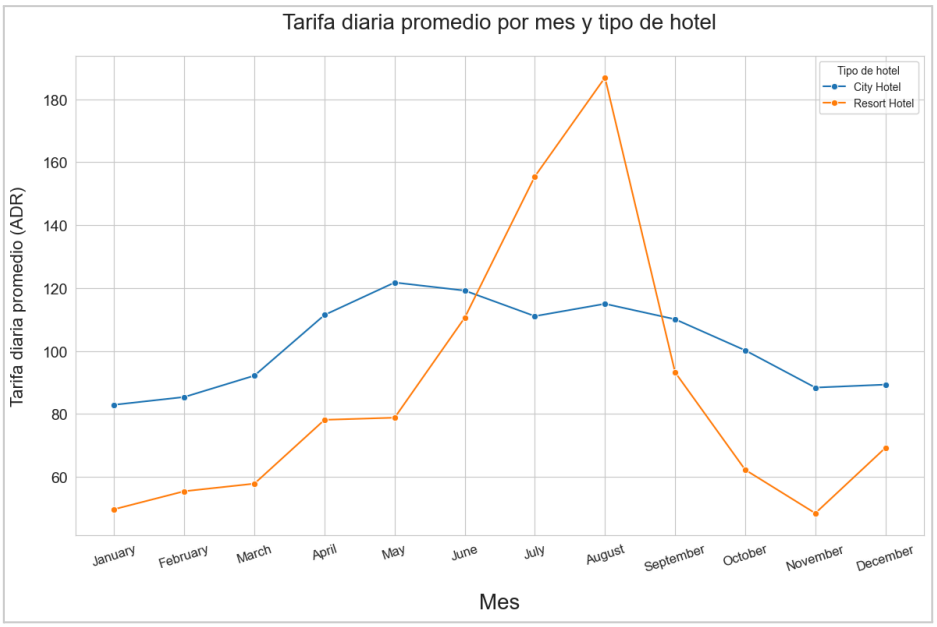


Dentro de la distribución de personas, podemos ver como por lo general los adultos suelen reservar en par, mientras que para sus acompañantes más jóvenes es común que no se suelan hacer reservaciones para niños o bebés. Ahora, sería interesante analizar la edad de las personas adultas que se suelen hospedar en los hoteles para poder ajustar las conclusiones y recomendaciones a esto, aunque es importante considerar las políticas de privacidad para los datos de los clientes que se hospedan.

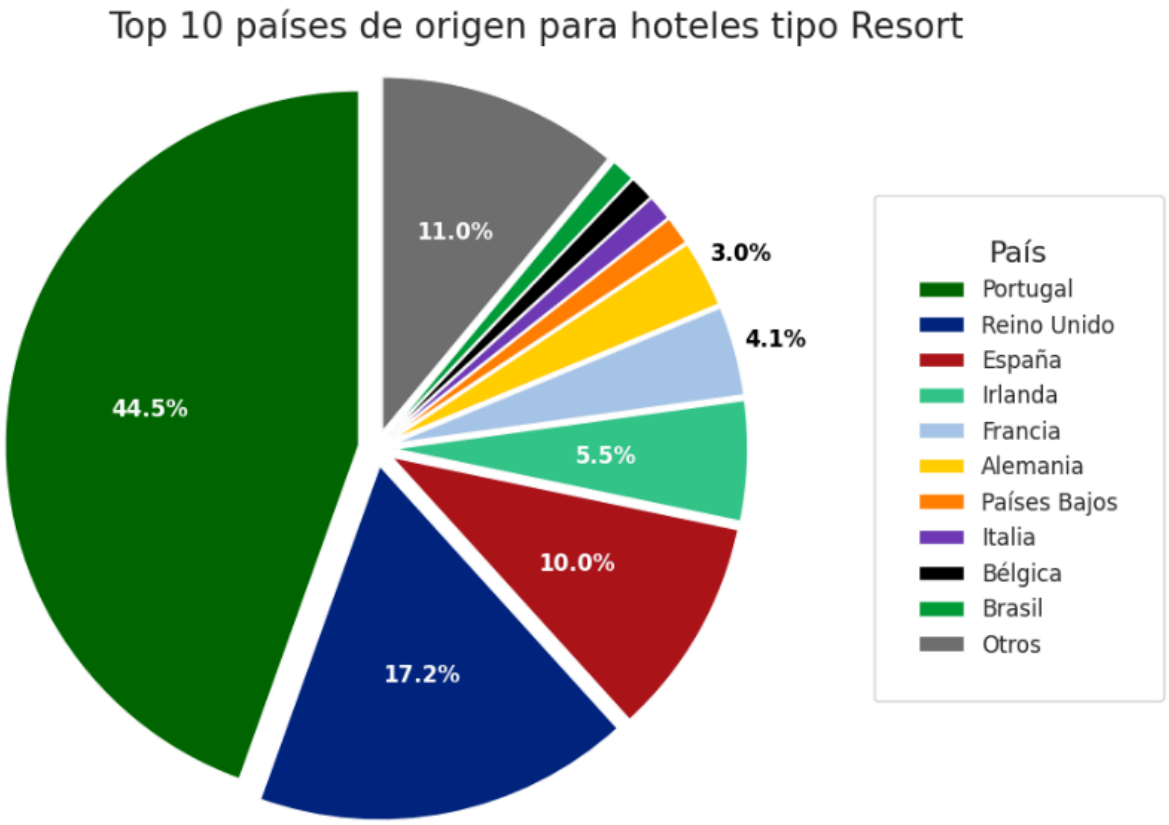


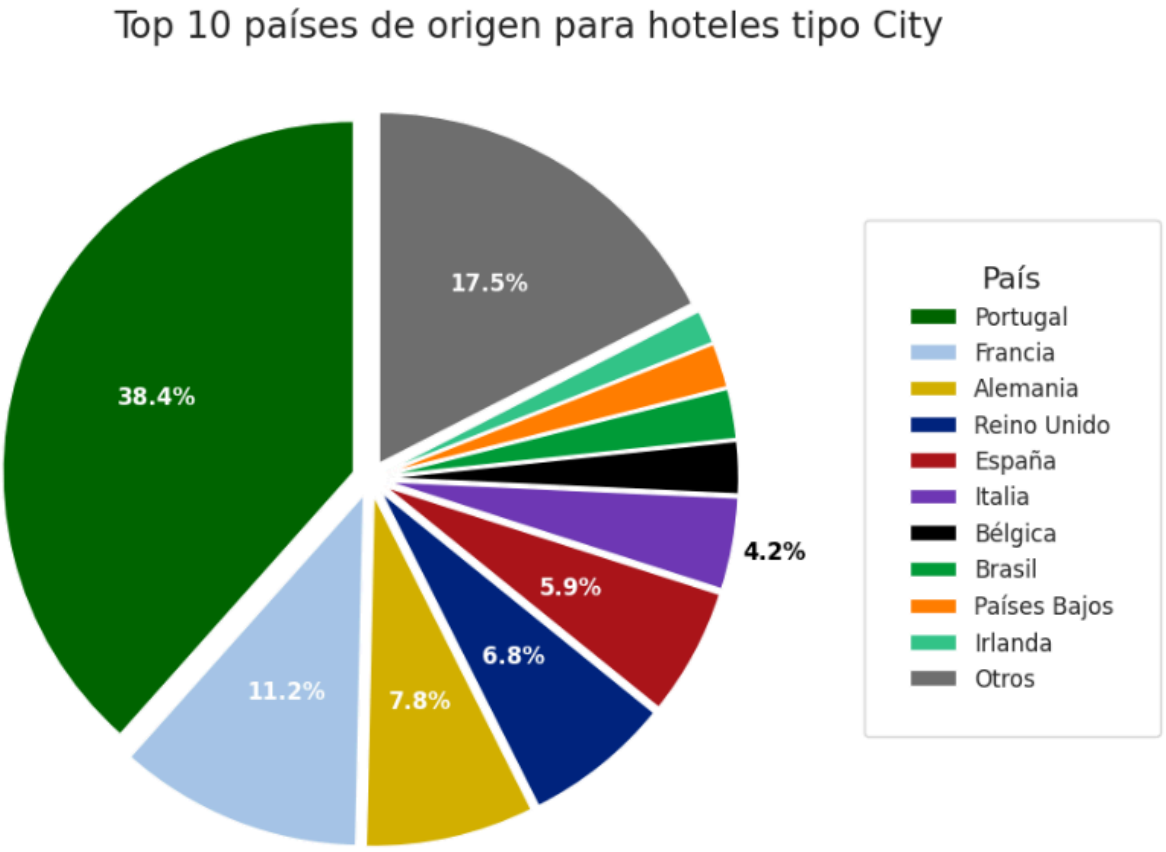
Para la distribución de ADR, podemos ver como existe una asimetría a la derecha, donde la mayoría de los datos se encuentra entre 50 y 100 de ADR. Esto nos da una mejor idea de los valores comunes que suele tomar ADR y enfocarnos en las tarifas más altas que pueden presentar una buena oportunidad de ingresos si se hacen los ajustes correctos.





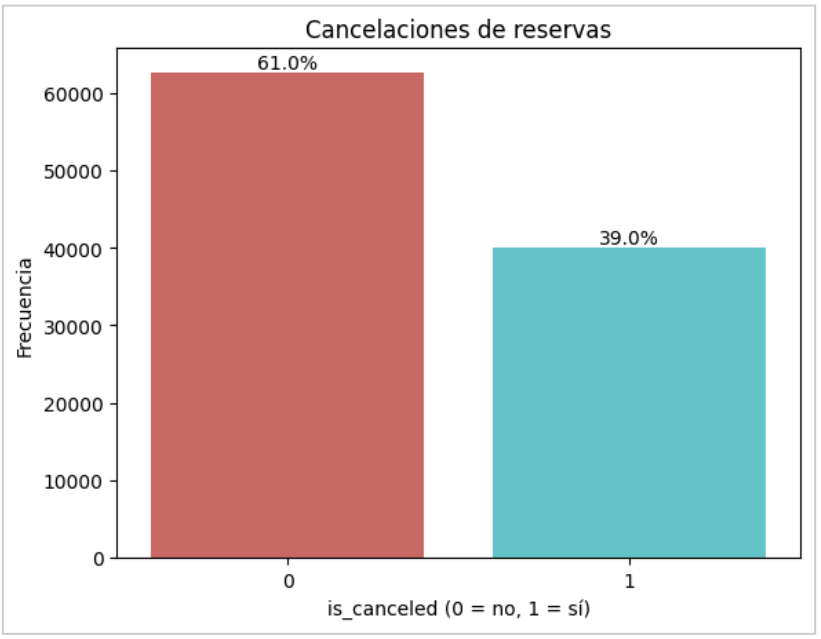
Al usar esta serie de tiempo, podemos ver de una mejor manera como se comportan las tarifas a lo largo de los años. Esto nos confirma (como lo habíamos visto antes) que en verano es donde suele alzarse la demanda, más específicamente para los de tipo Resort. Así también, es interesante ver como también existen picos hacía los días de año nuevo para ambos tipos de hotel, lo que representa una gran oportunidad a tener en cuenta.



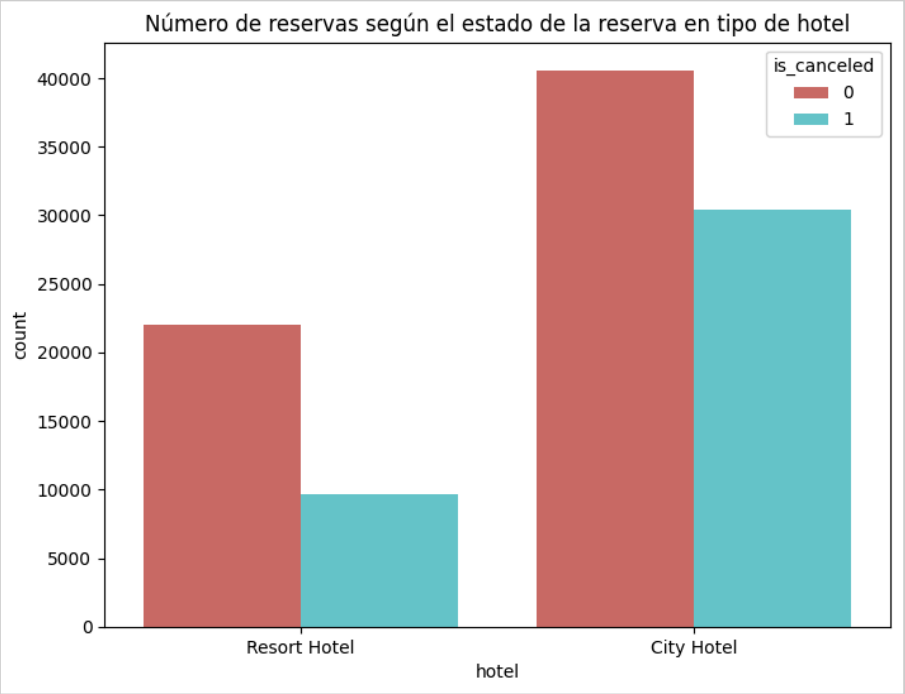


Para los países de origen de los huéspedes podemos observar cómo gran parte suele ser gente local, y algunos otros países como España, Francia y Reino Unido son países que también representan una parte considerable de los países de origen. Con esto podemos enfocarnos en estos segmentos de clientes a la hora de hacer proyectos de marketing y paquetes dirigidos a estos para poder atraer más gente y aumentar los ingresos.

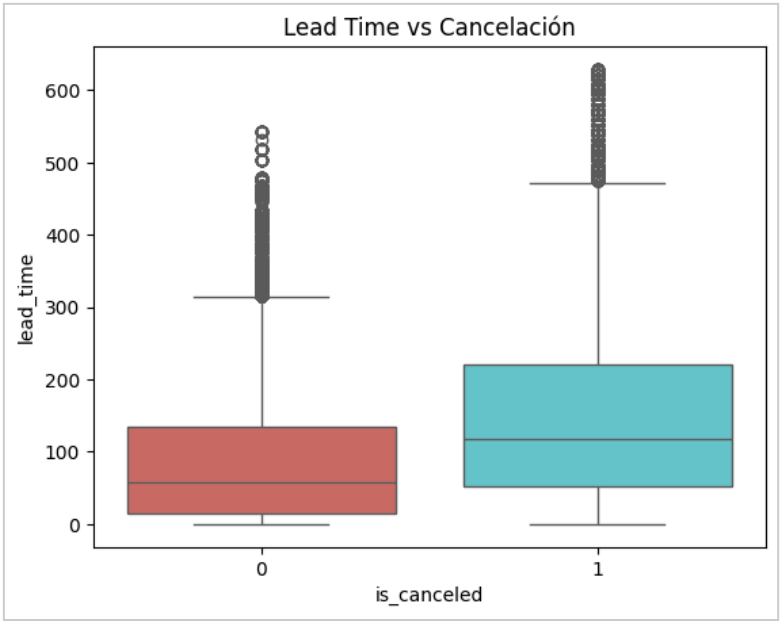
Objetivo 2: Reducción de cancelaciones



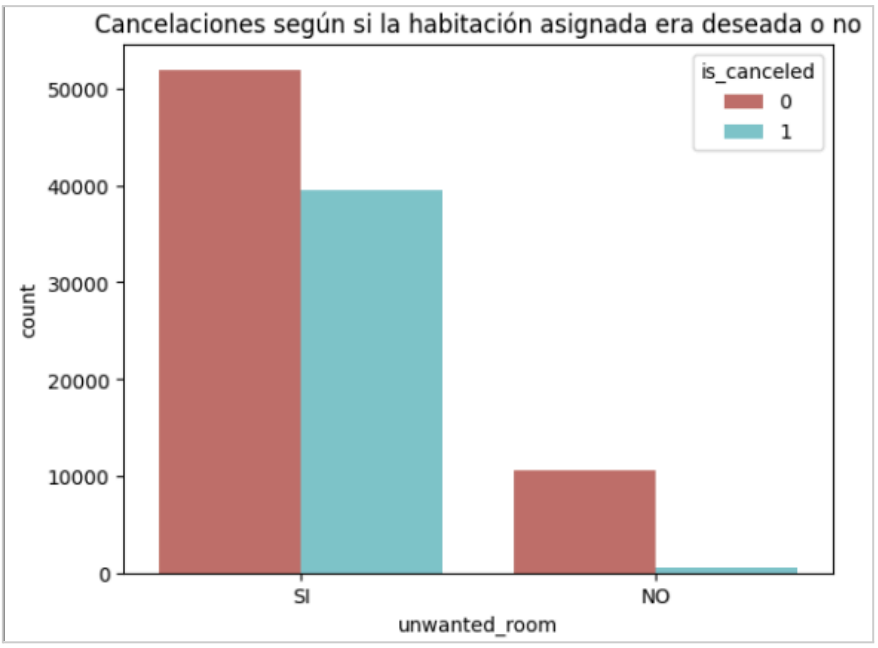
La gráfica anterior nos permite observar nuestra variable objetivo, y nos dice que no es tan dispareja, teniendo un 39% de cancelaciones, lo cual es un porcentaje considerablemente alto.



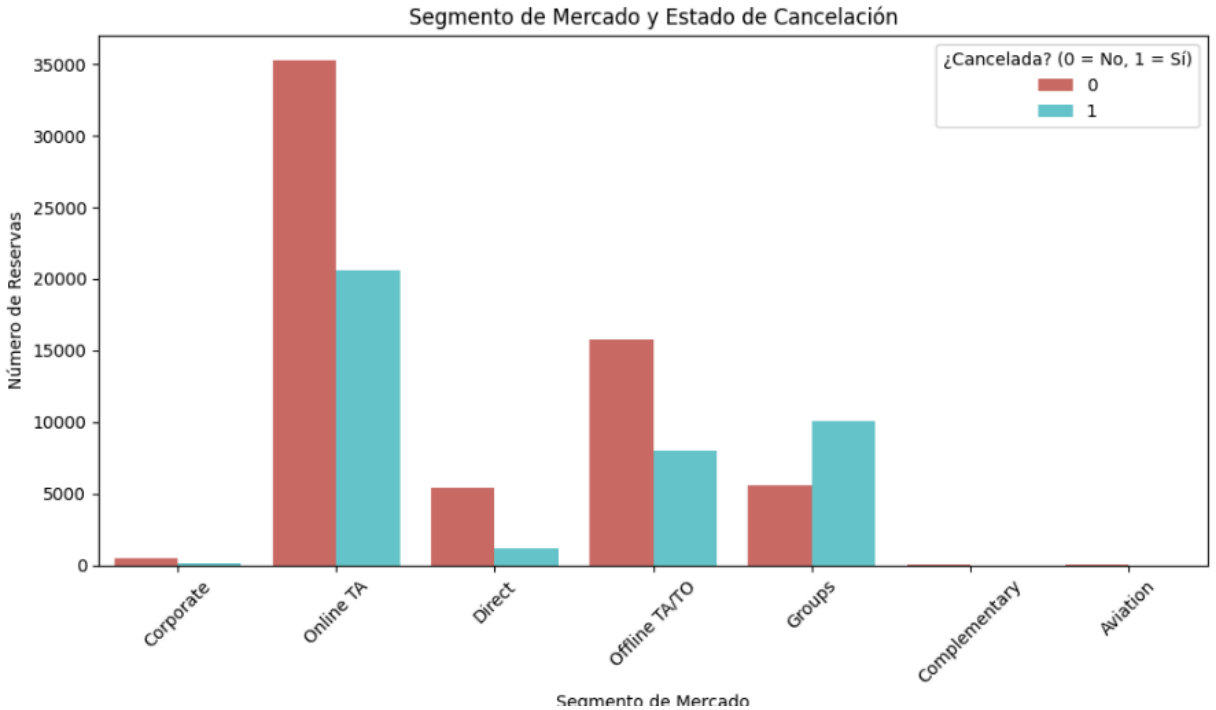
Aquí podemos ver que los hoteles en la ciudad tienen una mayor cantidad de reservas, pero al mismo tiempo la tasa de cancelación es considerablemente más alta. Teniendo por un lado una cancelación de alrededor de 30% en los hoteles tipo resort, mientras que en el lado de los hoteles de ciudad tenemos una tasa de cancelación de alrededor de 42%.



Graficando el boxplot entre la *Cancelación* y el *Lead Time* podemos ver que las reservas que se cancelan suelen hacerse con mucha mayor anticipación que las que no se cancelan, por lo que podemos decir que un lead time alto aumenta la probabilidad de cancelación.

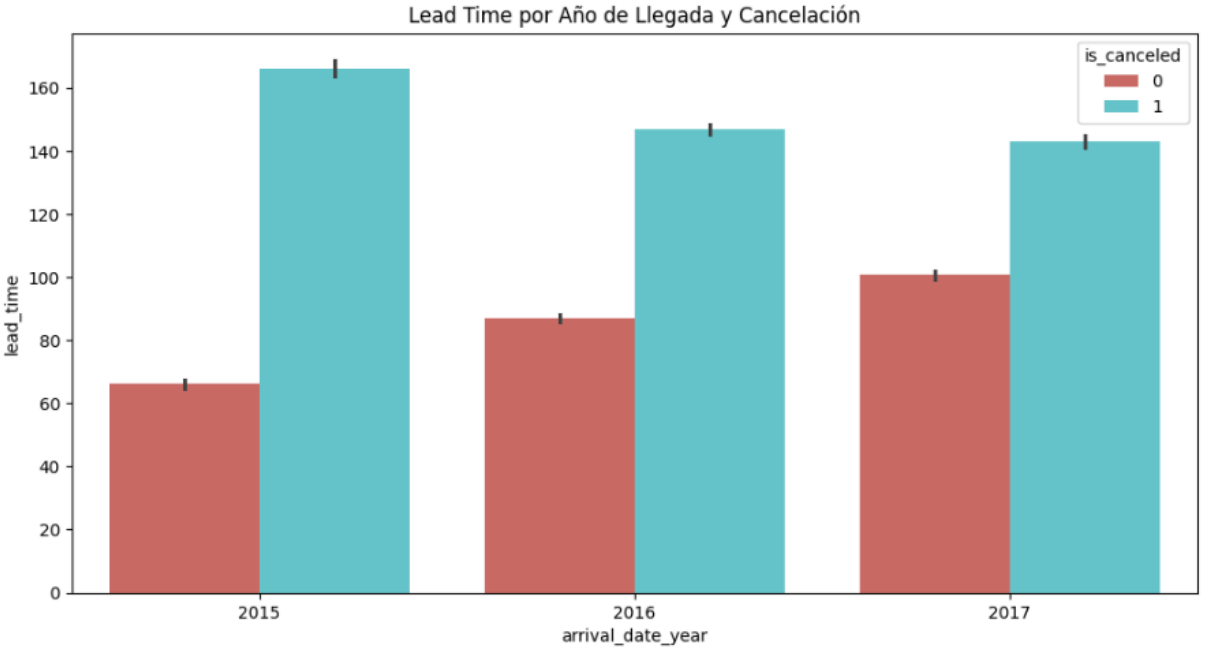


Al contrario de lo que uno podría imaginar, que una habitación asignada no coincida con el tipo de habitación solicitado no influye en la cancelación de la reserva, pues los datos muestran que la tasa de cancelación en estos casos es muy baja.

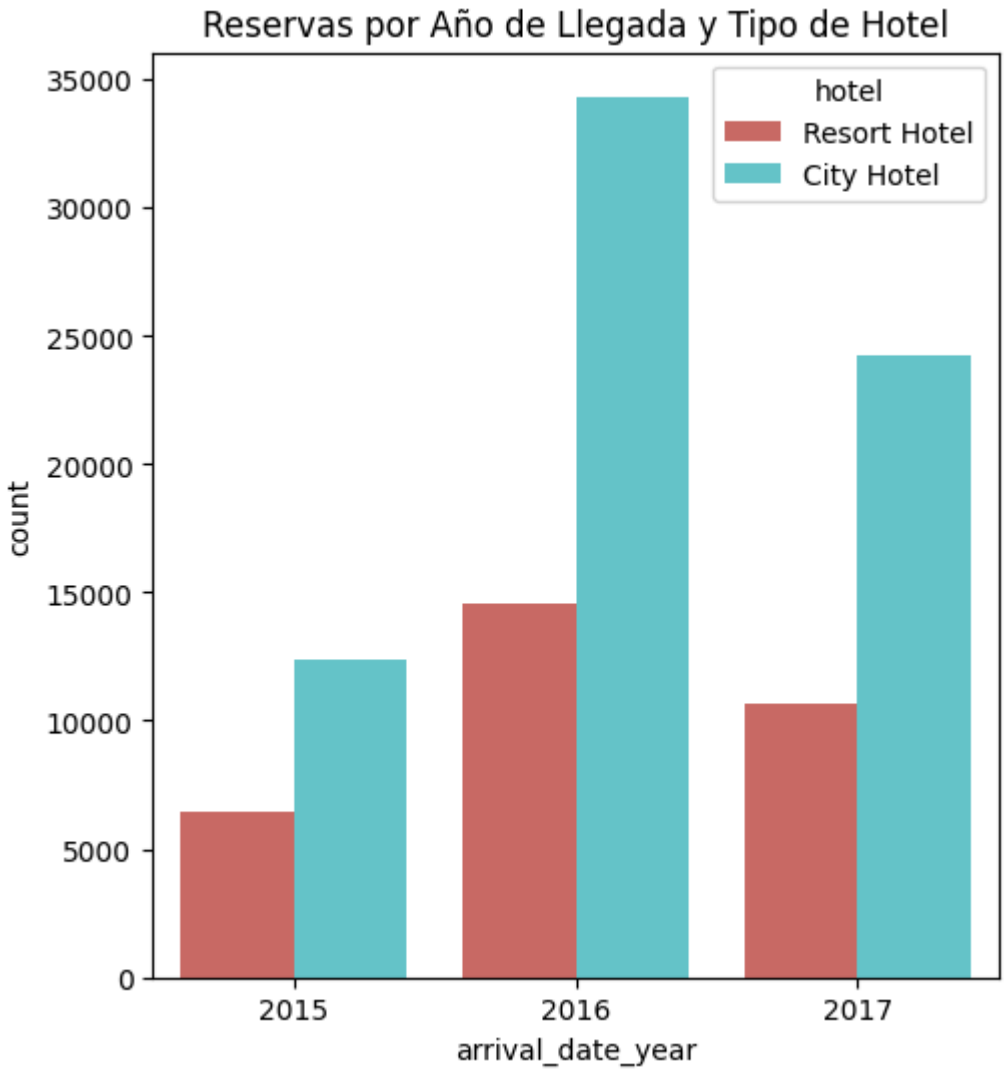


Para esta gráfica podemos ver que por lo general la proporción de cancelaciones es bastante considerable si la comparamos para reservas no canceladas. Es evidente como el canal en línea es muy popular para hacer las reservas de este tipo de hotel, pero al mismo tiempo presenta una gran proporción de cancelaciones lo que es necesario tener en cuenta, y respecto a las reservaciones de tipo group podemos ver son todavía mayor el número de cancelaciones ante la que no fueron canceladas. Dentro de estos dos aspectos creemos que es clave indagar porque pasa esto y establecer acciones que ayuden a incrementar el número de cancelaciones que se hacen para todos los canales en general (como incentivos, descuentos o hasta penalizaciones por cancelación si fuera necesario para esos casos), en especial para agencias

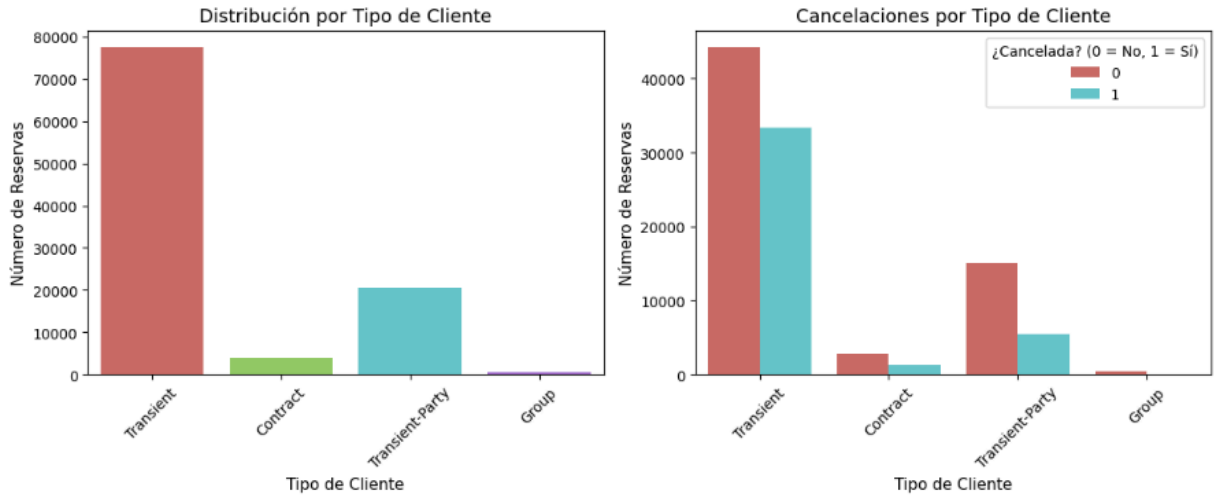
de viajes en línea que son las que más ingresos representan pero al mismo tiempo más cancelaciones presenta.



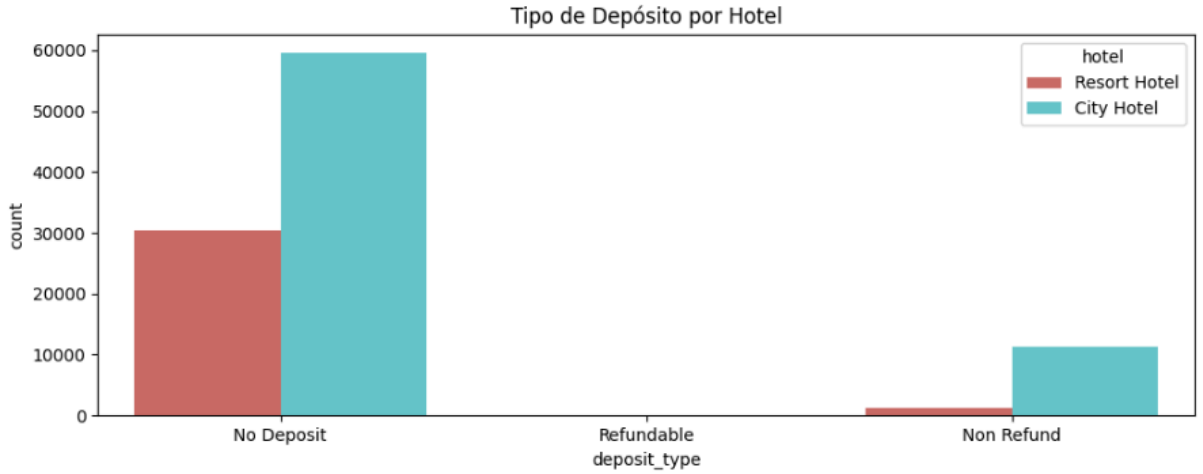
Esta gráfica nos da uno de los insights más importantes. La gráfica promedia el plazo de llegada (es decir, entre que se hace la reserva y el día del check in de manera hipotética en el caso de que se cancele) para los diferentes años que se tiene datos. Según la gráfica las cancelaciones se suelen hacer cuando el lead time es mayor, es decir, los clientes son más propensos a cancelar entre mayor sea el plazo entre la reservación y la llegada al hotel. Contextualmente tiene sentido, pues si los clientes reservan con mucho tiempo de antelación es más probable que tengan más imprevistos a lo largo del tiempo que pasa antes de la fecha prevista de llegada al hotel.

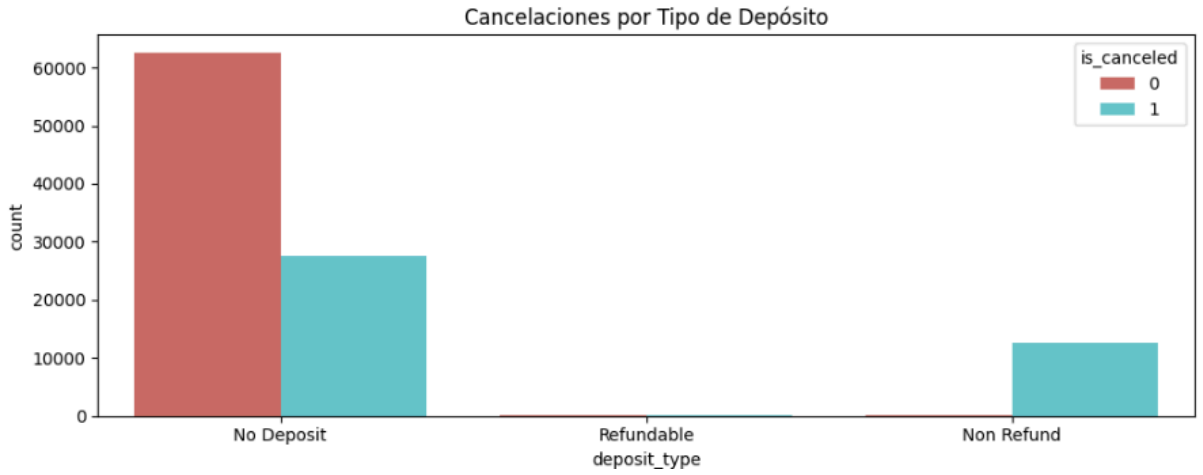


Está gráfica nos da un poco de mejor visualización de nuestros datos y nos da a entender que los datos más representativos son para el año 2016. Así también, confirma que hay más reservaciones para el hotel de tipo ciudad.

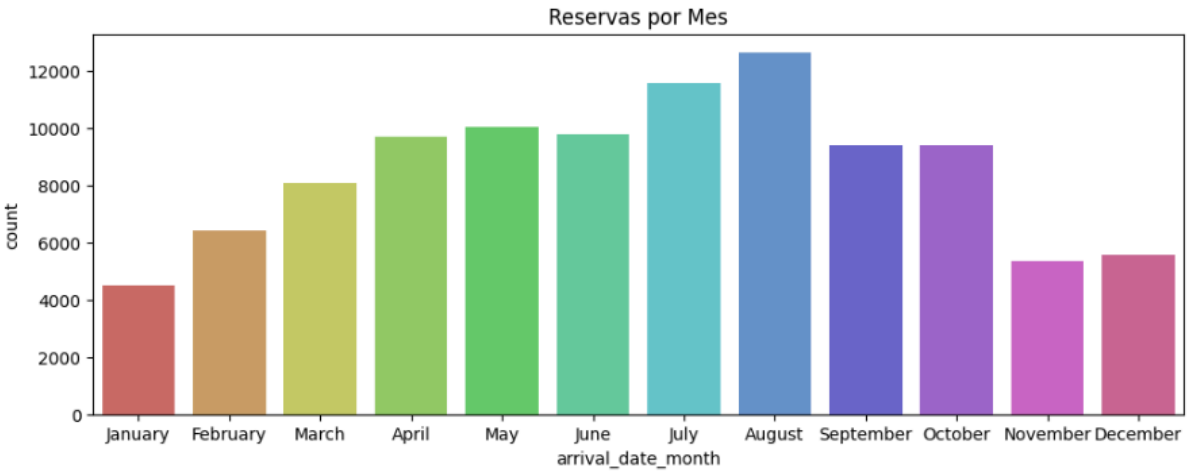


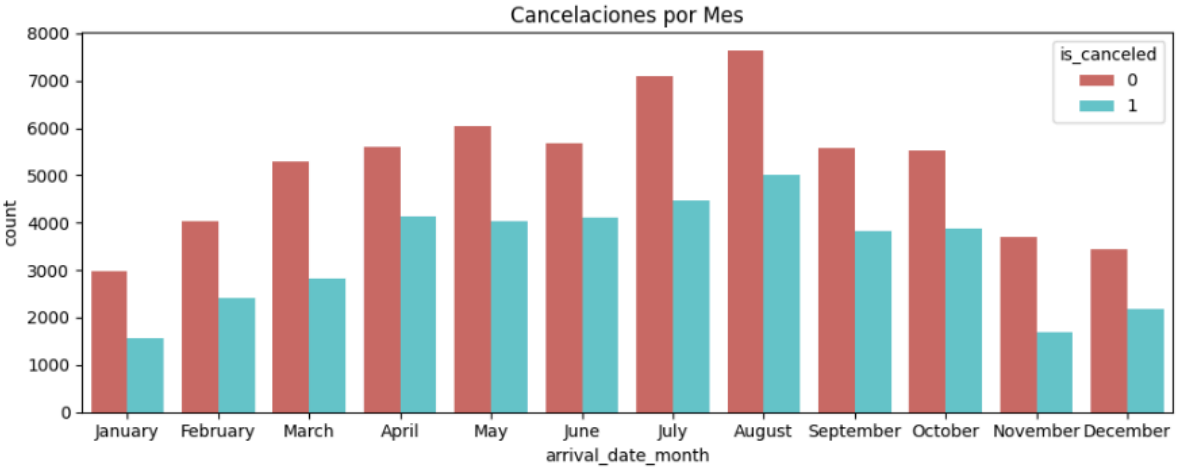
Para la comparación entre tipo de reserva/cliente, podemos ver que en su mayoría las reservas son hechas de manera individual (transient) y que este mismo tipo de reserva es en la que nos deberíamos enfocar a la hora de hacer nuestras recomendaciones y posterior plan de acción.



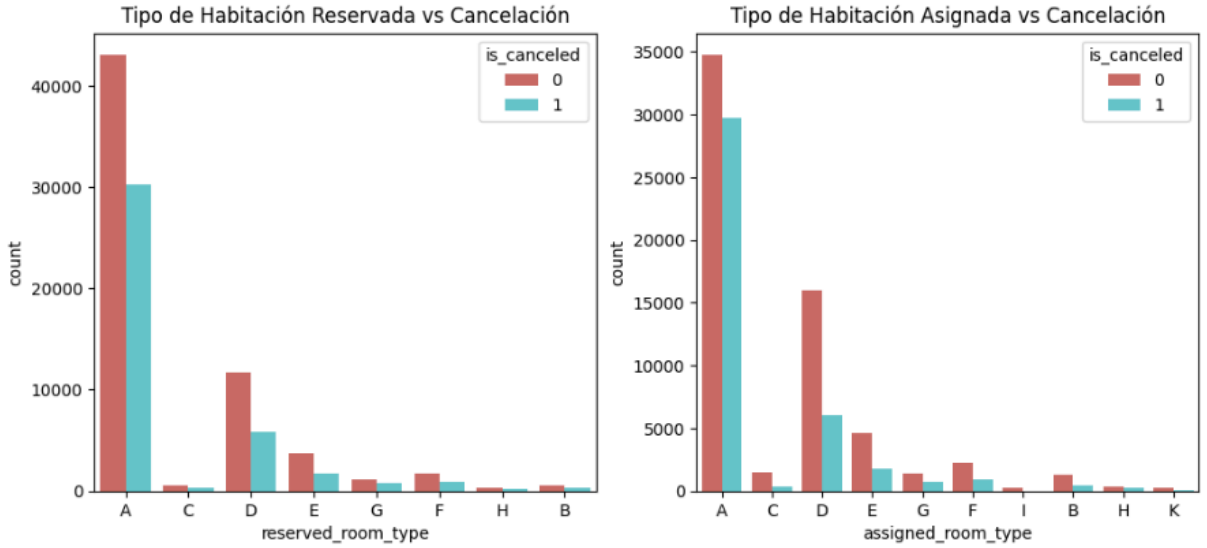


Ahora haciendo un análisis de estas dos gráficas, encontramos lo siguiente: aunque la modalidad "No Deposit" es la estándar y muestra un comportamiento de cancelación equilibrado, las reservas marcadas como "Non Refund" (No Reembolsables) presentan una tasa de cancelación muy alta. Esto resulta algo extraño, pues una penalización en este rubro debería ser un incentivo para no cancelar, lo que sugiere que podría haber algún tipo de sesgo en los datos o un factor externo para esto. Aquí también podemos ver que es muy popular en la industria no pedir ningún tipo de depósito al hacer la reservación.



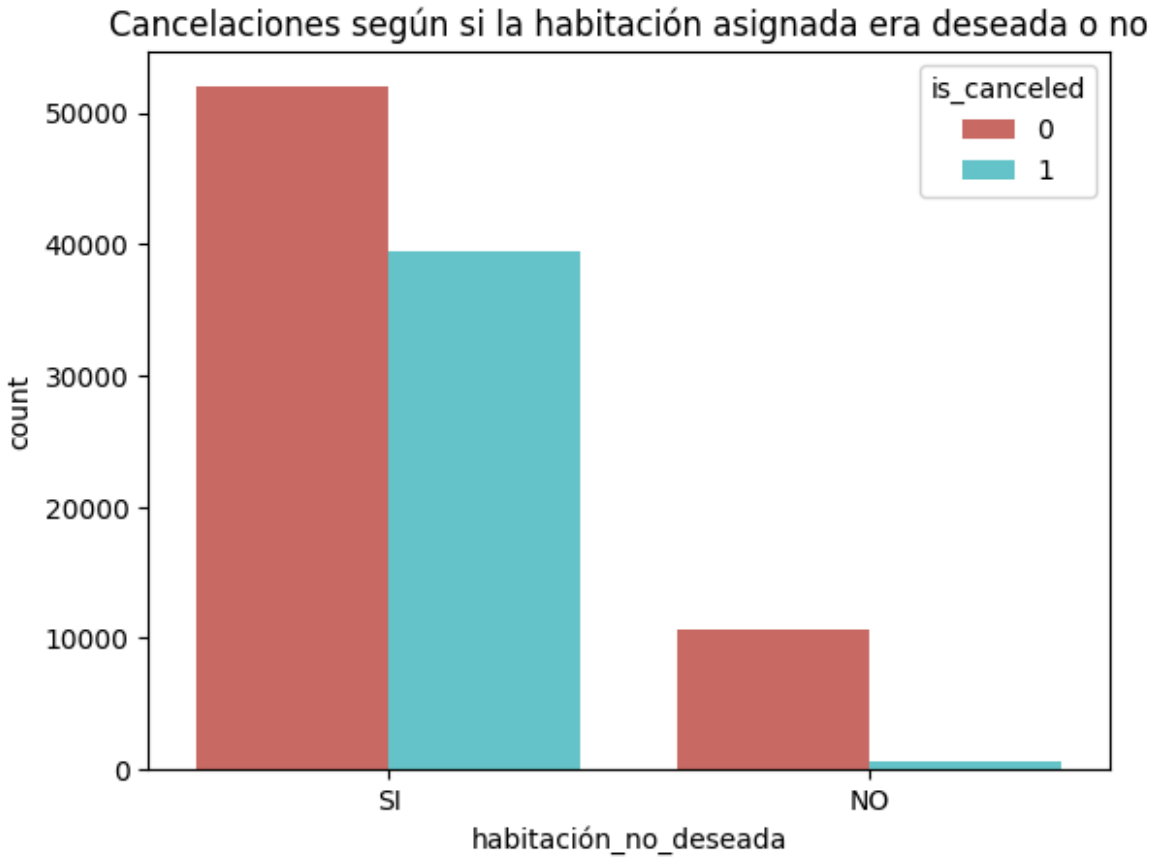


Estas dos gráficas reafirman lo que ya veníamos analizando anteriormente en el objetivo 1 acerca de la estacionalidad. Existe una demanda más alta para los meses de verano (agosto, julio y junio).



El análisis comparativo entre el tipo de habitación reservada y la asignada revela una algo clave para la gestión de ingresos: mientras que la Habitación Tipo A (la categoría estándar) concentra el mayor volumen de demanda y también el mayor riesgo de cancelación, se observa un cambio positivo en la distribución de las habitaciones asignadas. Las categorías superiores

de cuartos (como D y E) muestran una tasa de retención más alta en la asignación final que en la reserva inicial. Esto sugiere que la reasignación de habitaciones (como mejoras por falta de cuartos en la categoría A) actúa como un incentivo eficaz que reduce la intención de cancelación, pues se le entrega al cliente un valor mejor al esperado. Ahora, sería interesante tener más información de cada tipo de cuarto, pues en el dataset inicial fueron censurados este tipo de datos por anonimidad.



Con esto, se confirma que la gestión del inventario resulta muy importante en la retención de clientes. La gráfica muestra una diferencia importante en el comportamiento del consumidor: mientras que el grupo estándar (donde la habitación asignada coincide con la reservada) mantiene las tasas comunes de cancelación, los clientes que experimentaron un cambio de habitación presentaron una tasa de cancelación muy baja. Esto valida la hipótesis de que la percepción de valor añadido al recibir una habitación superior a la contratada evita y disminuye el riesgo de cancelación.

Modelo de entrenamiento

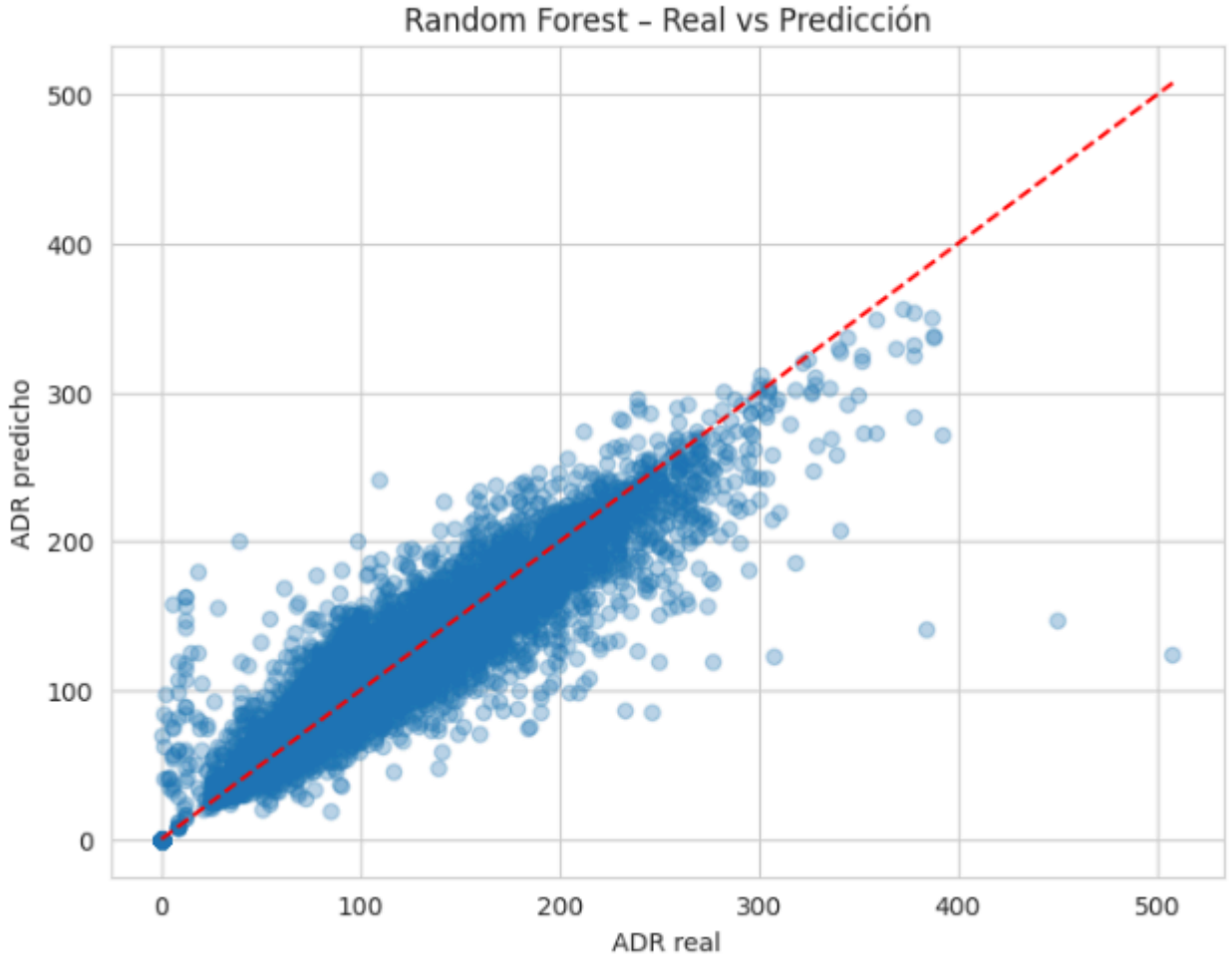
Objetivo 1 (Optimización de Ingresos)

A continuación, se presentan los features utilizados para nuestra etapa de entrenamiento en relación al objetivo uno, (Optimización de ingresos):

Los siguientes features fueron descartados del entrenamiento debido a que se consideró que no tenían utilidad para el objetivo de este proyecto:

- is_canceled
- agent
- assigned_room_type
- company
- booking_changes
- reservation_status
- reservations_status_sate
- required_car_parking_spaces
- total_of_special_requests

Para este caso se utilizó un Random Forest Regressor para hacer la predicción del ADR:



Cabe resaltar que en este caso desarrollamos un modelo de regresión y no de clasificación (como el de cancelaciones). Por lo tanto, aquí no se busca la probabilidad de pertenecer a una categoría, sino que, mediante los datos de entrenamiento, buscamos predecir un valor numérico continuo: el precio medio diario (ADR).

Al analizar los resultados, vemos que el desempeño es bastante sólido. Si bien persisten algunos outliers (por ejemplo, en rangos de 500 con errores grandes, o desplazamientos verticales en 0 y 100 que podrían deberse a descuentos o errores de captura que asumimos como válidos), la gran mayoría de las predicciones se mantienen dentro de un margen de error muy aceptable.

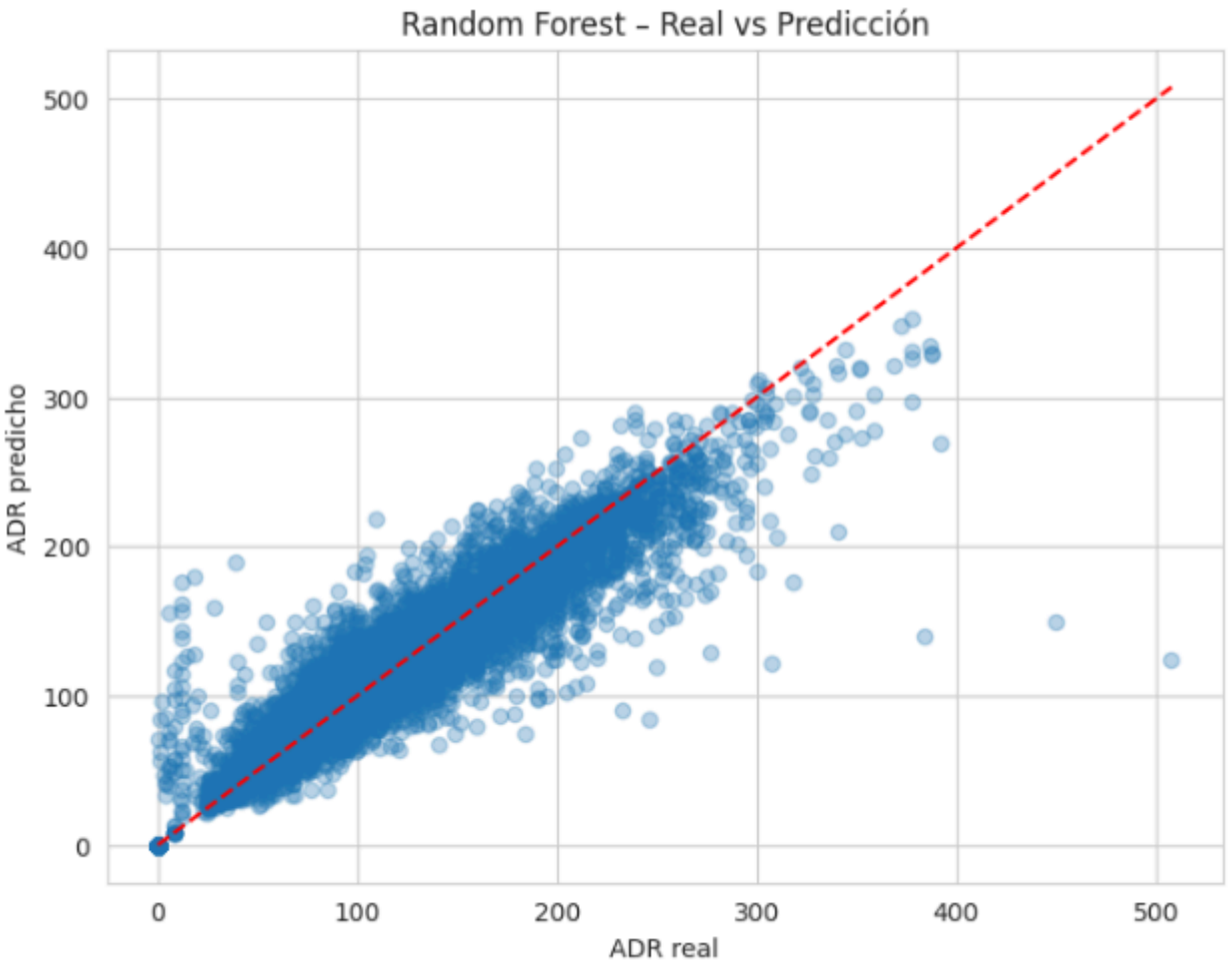
Analizando las métricas R2 y RMSE (enfocándonos en el conjunto de prueba/Test):

RMSE (14.43): Nos indica que el error promedio del modelo es de 14.43 unidades de dinero. Esto es una mejora considerable respecto a los 19 anteriores que se hizo en una prueba anterior.

R2 (0.910): Nos muestra que tenemos un 91% de explicabilidad de los datos en el conjunto de prueba (e incluso un 96.7% en entrenamiento). Esto significa que el modelo es capaz de explicar la inmensa mayoría de la variabilidad del precio.

Es un resultado excelente, considerando que se trata de un modelo base al cual todavía no se le ha aplicado un ajuste fino de hiperparámetros (tuning). Sin embargo, esto también podria indicar overfitting.

El objetivo de este ajuste fue evaluar y mejorar la capacidad de generalización del modelo. Al introducir estos parámetros, logramos mitigar el sobreajuste (overfitting), evitando que el modelo simplemente "memorice" los datos de entrenamiento. Si bien el error (RMSE) no disminuyó drásticamente, la relación entre las métricas de entrenamiento y las del conjunto de prueba (el 20% de datos no vistos) indica que el modelo es mejor y confiable ante nuevos datos.



En conclusión, esta versión del modelo demuestra una mejor capacidad de generalización en comparación con los intentos previos. Si bien se ha logrado reducir el sobreajuste (overfitting) , aún persisten ciertos márgenes de error en las predicciones. Esto sugiere que, aunque las variables actuales explican satisfactoriamente el ADR, el modelo se beneficiaría de enriquecer el dataset con atributos adicionales, como indicadores explícitos de descuentos o promociones estacionales, para contextualizar mejor los valores atípicos. Considerando el tamaño de la muestra (116k registros), este resultado representa el mejor balance obtenido hasta ahora entre precisión y generalización.

Respecto a la importancia de las características, la distribución de los pesos se ve influenciada por el gran número de variables obtenidas que resultó de la codificación de variables categóricas (más de 215 columnas tras el proceso de dummies). A pesar de esta dilución, resulta revelador que variables como la cantidad de niños, el plazo de llegada, el tipo de hotel (Resort) y el número de adultos destaquen como los predictores más influyentes en la determinación del precio.

Objetivo 2 (Cancelaciones)

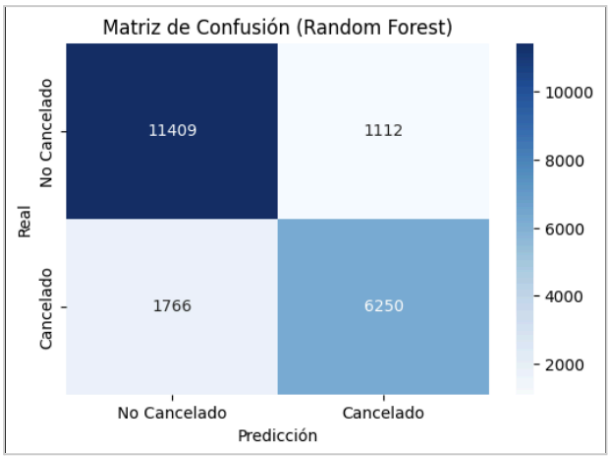
Los siguientes features fueron descartados para el entrenamiento debido a que no tenían utilidad y además eran columnas que contenían información posterior al momento de la reserva para prevenir data leakage.

Variables descartadas

- reservation_status
- reservation_status_date
- assigned_room_type
- booking_changes
- days_in_waiting_list
- arrival_date_year
- arrival_date_month
- arrival_date_week_number
- arrival_date_day_of_month

Con el objetivo de poder predecir la probabilidad de una cancelación de alguna reserva (is_canceled) utilizamos dos modelos, Random Forest Classifier y XGBoost Classifier, ya que contamos con una gran cantidad de variables y también con conjuntos de datos con múltiples variables categóricas las cuales transformamos.

El modelo de Random Forest lo entrenamos con las variables disponibles tras la limpieza de datos, usando 400 árboles y una semilla fija en donde 80% de los datos se utilizaron para el entrenamiento mientras que el otro 20% para prueba.




```
Accuracy: 0.8598626868578663
[[11409  1112]
 [ 1766  6250]]
              precision    recall  f1-score   support

           0       0.87      0.91      0.89     12521
           1       0.85      0.78      0.81      8016

    accuracy                           0.86     20537
   macro avg       0.86      0.85      0.85     20537
weighted avg       0.86      0.86      0.86     20537
```

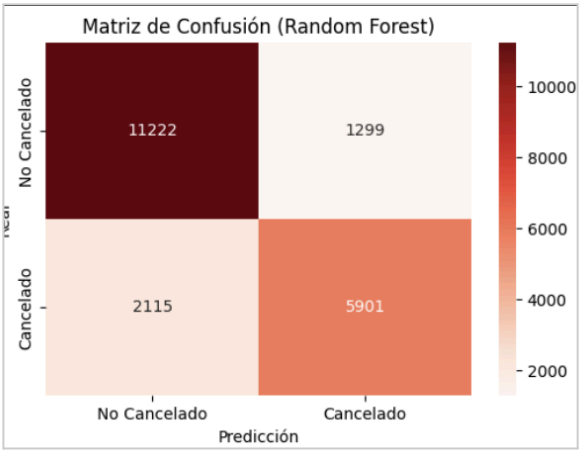
Obtuvimos que el modelo tenía una precisión del 87% y un recall de 0.9 para la clase de cancelaciones, lo que indica un buen desempeño general pero con presencia de falsos negativos.

Intentamos mejorar el modelo mediante selección de variables con mayor correlación con el objetivo, y un ajuste de parámetros pero el desempeño del modelo se redujo, por lo que consideramos al modelo original como el más adecuado para estimar el riesgo de cancelación.

```
Accuracy: 0.8337634513317427
[[11222  1299]
 [ 2115  5901]]
              precision    recall  f1-score   support

           0       0.84      0.90      0.87     12521
           1       0.82      0.74      0.78      8016

    accuracy                           0.83     20537
   macro avg       0.83      0.82      0.82     20537
weighted avg       0.83      0.83      0.83     20537
```

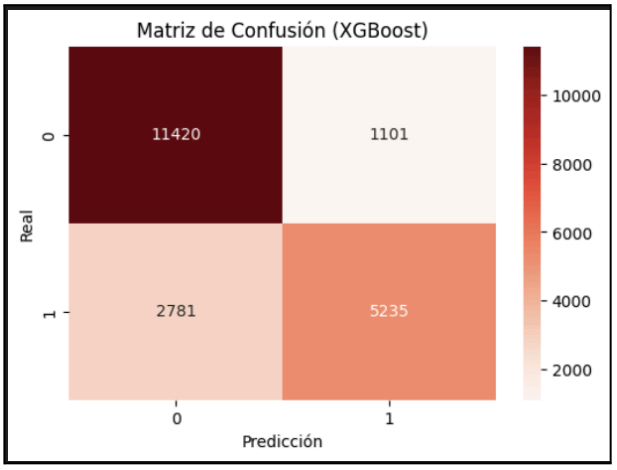


XG Boost

```
Accuracy: 0.8117188768057133
[[17163  1618]
 [ 4182  7842]]
              precision    recall  f1-score   support

           0       0.80      0.91      0.86     18781
           1       0.83      0.65      0.73     12024

    accuracy                           0.81     30805
   macro avg       0.82      0.78      0.79     30805
weighted avg       0.81      0.81      0.81     30805
```

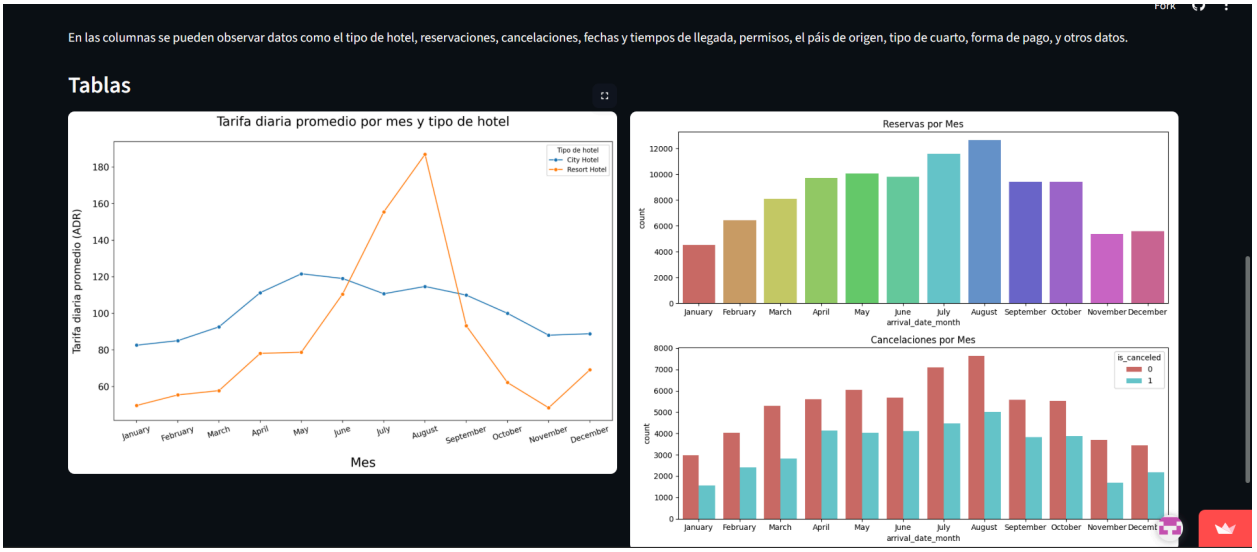


El modelo XGBoost logra un buen desempeño genera, alcanzó un accuracy cercano al 81% con una precisión de 0.8,indicándonos un buen desempeño en general, el cual clasifica correctamente la mayoría de las reservas que no se cancelan, pero reduce más el recall para cancelaciones reales en comparación con Random Forest. Esto se refleja en un número mayor de falsos negativos (2781).

Streamlit

https://bi-dash-2025.streamlit.app/

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	children	ba
0	Resort Hotel	0	342	2015	July	27	1	0	0	2	0	
1	Resort Hotel	0	737	2015	July	27	1	0	0	2	0	
2	Resort Hotel	0	7	2015	July	27	1	0	1	1	0	
3	Resort Hotel	0	13	2015	July	27	1	0	1	1	0	
4	Resort Hotel	0	14	2015	July	27	1	0	2	2	0	
5	Resort Hotel	0	14	2015	July	27	1	0	2	2	0	
6	Resort Hotel	0	0	2015	July	27	1	0	2	2	0	
7	Resort Hotel	0	9	2015	July	27	1	0	2	2	0	
8	Resort Hotel	1	85	2015	July	27	1	0	3	2	0	



Introducción | Cancelaciones | Optimización de Ingresos

❌ Cancelaciones

Modelo Random Forest

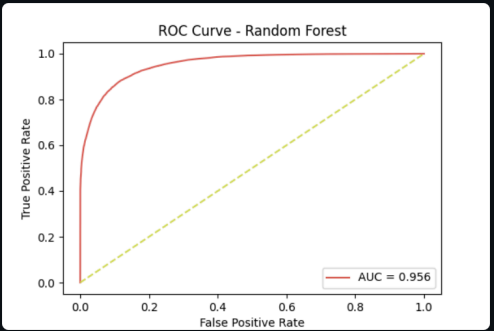
Predice que no cancelan

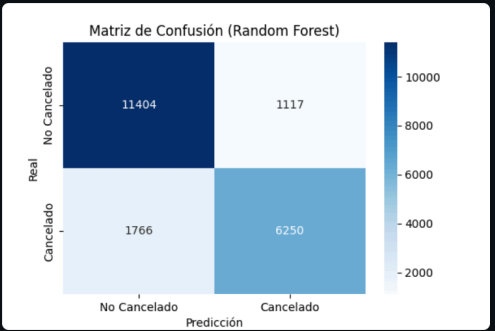
Precisión	Recall	f1_score
84%	88%	86%

Predice que cancelan

Precisión	Recall	f1_score
81%	79%	77%

Fork





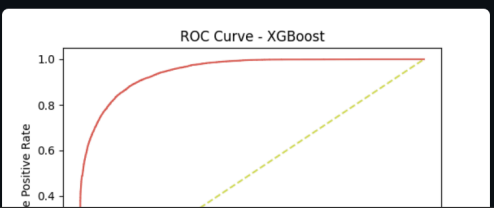
Modelo XGBoost

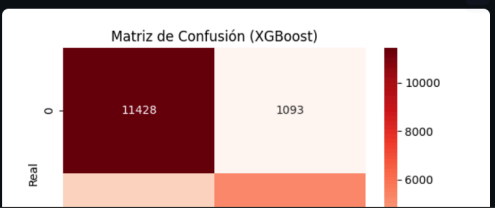
Predice que no cancelan

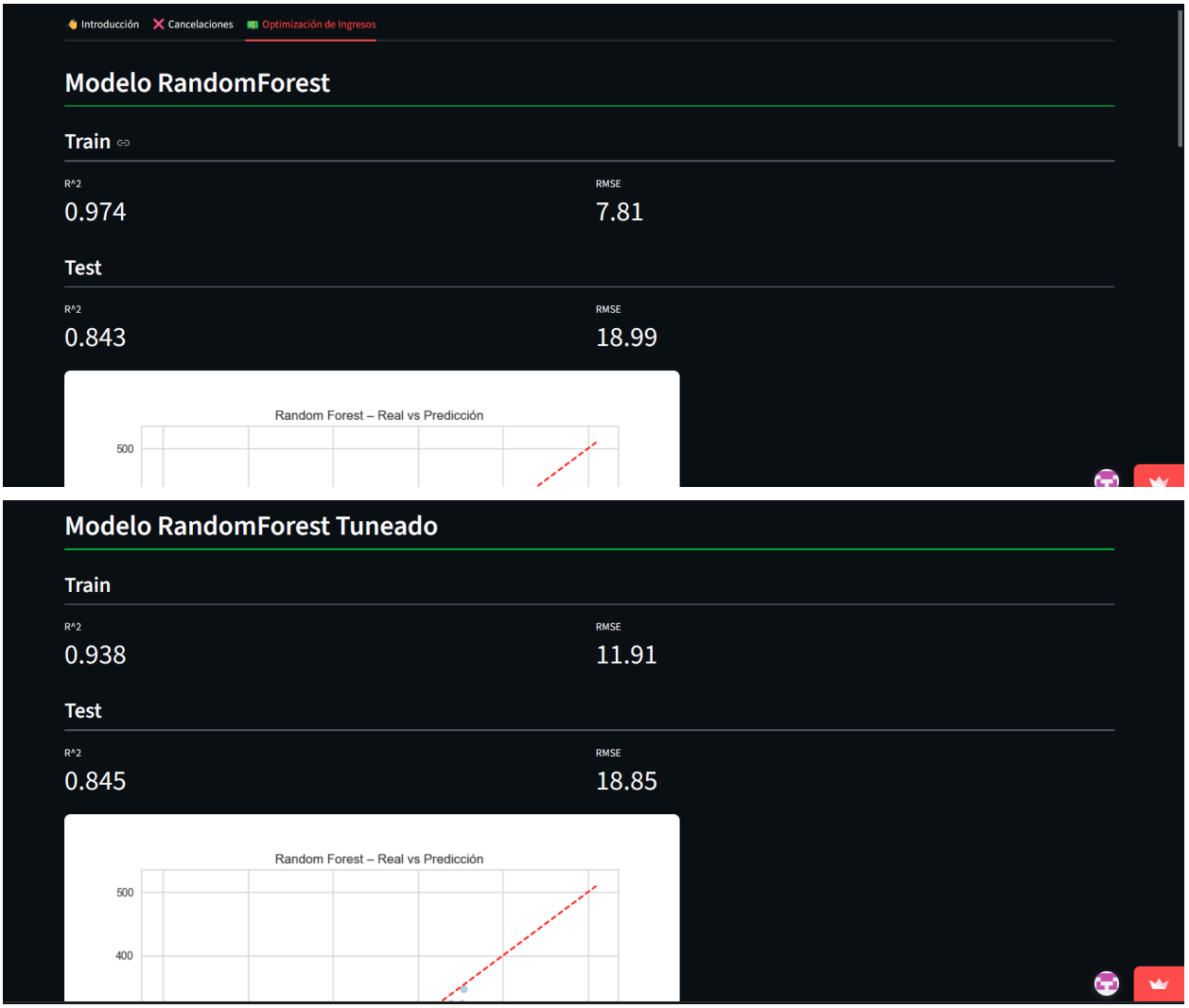
Precisión	Recall	f1_score
89%	90%	89%

Predice que cancelan

Precisión	Recall	f1_score
85%	82%	83%







Resultados y conclusiones

Primero, para la predicción de tarifas (ADR), el modelo de regresión Random Forest demostró un desempeño bastante alto, alcanzando una explicabilidad del 91% (R2) en el conjunto de prueba y un error promedio (RMSE) de 14.43 unidades.

El análisis gráfico fue de mucha ayuda y nos ayudó a identificar los siguientes puntos y poder realizar una buena estrategia de precios:

- Se detectó un comportamiento un tanto diferente entre los tipos de hotel relacionado a la estacionalidad. Mientras que los hoteles tipo Resort muestran un pico alto de tarifas en agosto (impulsado por la demanda de verano), los hoteles de ciudad mantienen tarifas estables sin fluctuaciones drásticas durante el año.
- Las variables demográficas (número de adultos y niños) resultaron ser las variables predictoras más influyentes, seguidas por el tipo de habitación. Esto último nos lleva a preguntarnos e indagar un poco más acerca del tipo de habitación real para las reservaciones, puesto que no tenemos información más específica sobre tipos de habitación por temas de anonimidad.

- Esto sugiere que la optimización de ingresos debe ser mixta, se deben aplicar precios dinámicos agresivos en verano para Resorts y enfocar los hoteles urbanos en mantener una ocupación constante.
- Así también, la segmentación del tipo de clientes es clave para aplicar estrategias de marketing enfocadas a ciertos sectores de clientes. Por ejemplo, para los hoteles tipo Resort, una gran parte es de origen local, es decir, portugués. En cambio, existe una relación más diversa en clientes para los hoteles de tipo ciudad, por lo que aquí la tarea sería la de desarrollar estrategias o paquetes que atraigan más a este tipo de clientes.
- Variables como el porcentaje de ocupación y la edad de los huéspedes podrían agregar mucho valor al análisis y a la estrategia posterior de aumento de ingresos, aunque esto podría ser complejo por temas de protección de privacidad de los clientes.

Para el tema de cancelaciones, el modelo Random Forest Classifier alcanzó una exactitud del 86% y un Recall del 91%, lo que nos permite detectar de manera eficaz a la gran mayoría de los clientes con riesgo de fuga.

El análisis de las variables reveló patrones de comportamiento críticos para la toma de decisiones:

- Pudimos confirmar mediante los diagramas de caja que el *lead_time* es la variable de mayor peso: a mayor tiempo de anticipación, mayor es la probabilidad de cancelación. Nosotros consideramos que una recomendación para los hoteles es que deberían desincentivar reservas a muy largo plazo sin garantías o lanzar promociones para reservas con ventanas más cortas, esto como alternativa a penalizar a los clientes por cancelar ya que en vez de evitar cancelaciones haría que se desincentivara las reservas antes que nada.
- Aunque las reservas sin depósito previo son las más comunes y propensas a cancelarse por la falta de compromiso financiero, se encontró una anomalía que llamaba la atención en las reservas "Non Refund", las cuales mostraron tasas de cancelación demasiado altas. Esto de cierta manera contradice la lógica que teníamos sobre las reservaciones y sugiere problemas en la captura del dato o factores externos que impactan nuestros resultados y se deben indagar a mayor profundidad con los hoteles.
- Un hallazgo positivo fue que cuando la habitación asignada es diferente de la reservada (generalmente mejoras del cuarto debido a fallas operativas), la tasa de cancelación cae drásticamente. Esto valida la reasignación de habitaciones como una poderosa herramienta de fidelización.
- También se observó que las reservas provenientes de agencias en línea (Online TA) aportan volumen pero traen consigo una tasa de cancelación considerablemente alta en

comparación con reservas directas. Por esto mismo es importante revisar este tipo de canales para ver si es posible colaborar con estas agencias y ofrecer descuentos o paquetes exclusivos por agendar por este medio, ya que la mayoría de las reservaciones (y por ende cancelaciones) se realizan por este.

Plan de Acción a Implementar

- Se debe implementar una tarifa dinámica basada en la estacionalidad, la demanda y el comportamiento histórico del ADR, poniendo especial atención en los meses de alta demanda, como agosto y julio, para maximizar los ingresos cuando la disposición a pagar es mayor.
- Es fundamental diferenciar la estrategia de precios según el tipo de hotel y tipo de habitación, debido a que existen diferencias significativas en el ADR entre ambos segmentos; esto nos indica diferentes niveles de disposición a pagar por parte de los clientes, siendo bastante mayor y más variable para los hoteles tipo Resort que para los de ciudad.
- Se recomienda priorizar las campañas de marketing y estrategias comerciales en los principales países de origen identificados (como España, Reino Unido y Francia), y muy especialmente en el mercado local (Portugal), ya que estos concentran la mayor proporción de reservas y representan el mayor impacto acumulado en los ingresos del negocio.
- Para mitigar las cancelaciones, los hoteles deberían establecer reglas de depósito y gestionar el *lead time*, ya que se demostró que las reservas realizadas con mucha anticipación y aquellas sin ningún depósito previo ("No Deposit") son las más propensas a cancelarse. Esto nos lleva a pensar en estrategias como implementar un depósito obligatorio para ciertos tipo de reserva o lanzar promociones para reservas con menor ventana de anticipación, lo que ayudaría a reducir este riesgo.
- Así también, se sugiere utilizar la asignación de habitaciones como estrategia de retención, dado que el análisis reveló que cuando la habitación asignada difiere de la reservada (generalmente mejoras), la tasa de cancelación disminuye drásticamente; esto valida el uso de mejoras de cuarto como incentivo para asegurar la estancia del cliente.
- Por último y como sugerencia, sería interesante recopilar información adicional para enriquecer el modelo, tal como el porcentaje de ocupación del hotel en tiempo real, el tipo específico de cuarto y la ubicación exacta de los hoteles, lo cual permitiría tener una visión más completa y corregir los márgenes de error actuales para futuras predicciones.