FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico 1

[75.06/95.58] Organizacion de Datos
Curso 1
Primer cuatrimestre de 2024

Alumno:	Roldán, Cristian	Nicolas Rizzo
Número de padrón:	96713	109756
Email:	croldan@fi.uba.ar	nrizzo@fi.uba.ar

Índice

1. Introducción

El presente informe documenta la solución del primer trabajo práctico de la materia Organización de Datos. Se realiza un análisis exploratorio de un conjunto de datos utilizando la biblioteca Pandas de Python. El objetivo principal es comprender la estructura del conjunto de datos, identificar patrones y características relevantes, y obtener información útil para su posterior análisis.

2. Metodologia

1. **Carga del conjunto de datos:** Se utilizó la función `read_csv()` de Pandas para leer el archivo CSV y cargarlo en un DataFrame.
2. **Exploración inicial:** Se examinaron las primeras filas del DataFrame para obtener una vista general de los datos.
3. **Información del DataFrame:** Se utilizó la función `info()` para obtener información detallada sobre el DataFrame, incluyendo el tipo de dato de cada columna, la cantidad de datos no nulos y las estadísticas descriptivas básicas.
4. **Preguntas de interes:** En el desarrollo del trabajo plantearemos preguntas y las trataremos de responder de acuerdo a los datos disponibles

3. Depuracion

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt


airline = pd.read_csv("./customer_airways_data.csv", encoding="iso-8
reviews = pd.read_csv("./cleaned-reviews.csv")
```

3.1. customer_airways_data.csv

Características del conjunto de datos

- Columnas categóricas:
  - Canal de venta: Internet o Mobile
  - Tipo de viaje: Ida, por tramos o ida y vuelta
  - Día de la semana: Abreviatura de los 7 días
  - Destino: Código IATA
  - Origen del país de la reserva
  - Preferencias binarias: Equipaje, comida y asiento
- Columnas numericas:
  - Número de pasajeros
  - Días entre la reserva y el viaje
  - Horario de vuelo
  - Duración del vuelo

Preprocesamiento

- No encontramos datos faltantes
- Se eliminaron 719 filas duplicadas, quedando un total de 4928 filas:
- Analisamos los paises que aparecen en bookin_origin: Encontramos algo particular, un dato de booking origin "(not set)". Eliminamos todas las filas correspondients a ese dato, ya que son 78 filas, correspondiendo a un 1.58 % del total.
- Se analizaron los valores atípicos, sin encontrar valores desproporcionados o errores. Se observaron algunos valores inusuales, como 9 pasajeros en algunas reservas, tiempos de vuelo superiores a 500 días y tiempos de estadía en destinos mayores a 400 días.

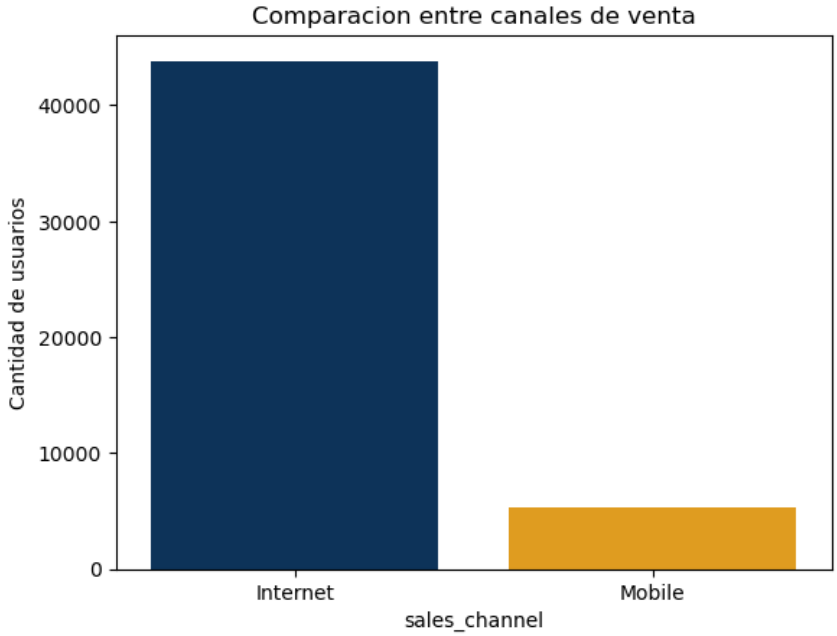
3.2. cleaned-reviews.csv

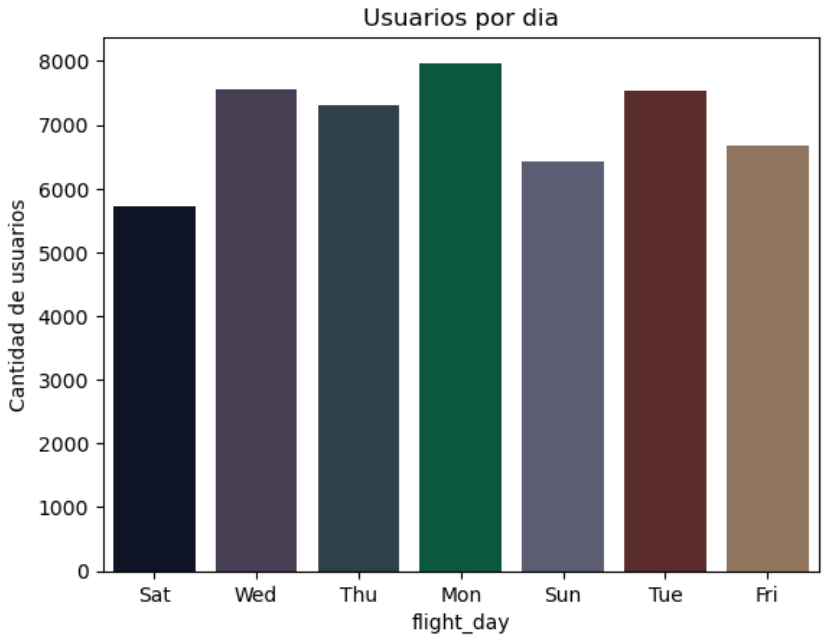
- Columnas categoricas:
  - Reviews
  - country: Entendemos que corresponde al lugar del review, no nacionalidad
  - verified: True para las review verificadas
  - comments: COmentarios de cada cliente sobre el servcio
- Columnas numericas
  - rate: Puntaje otorgado
  - unnamed:0 : Un indice que sospechamos no usar

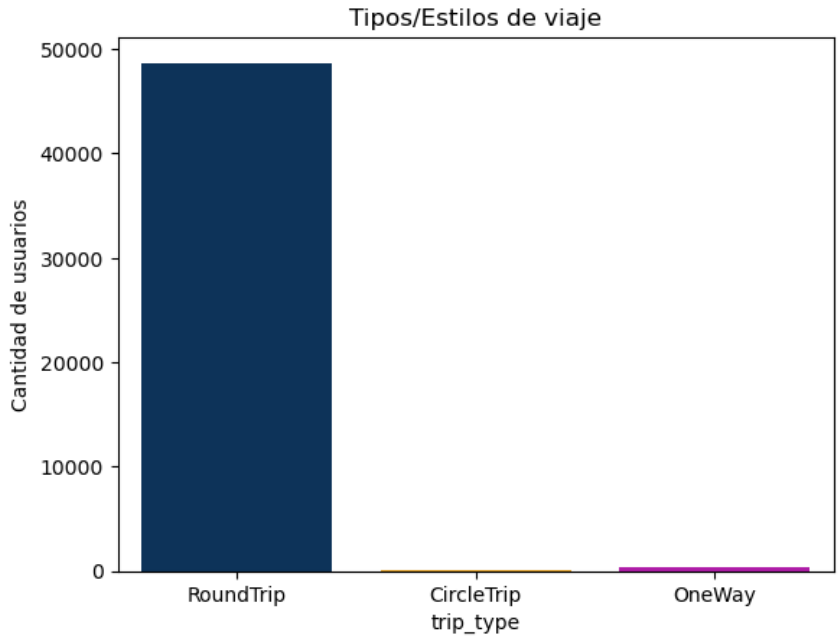
Preprocesamiento

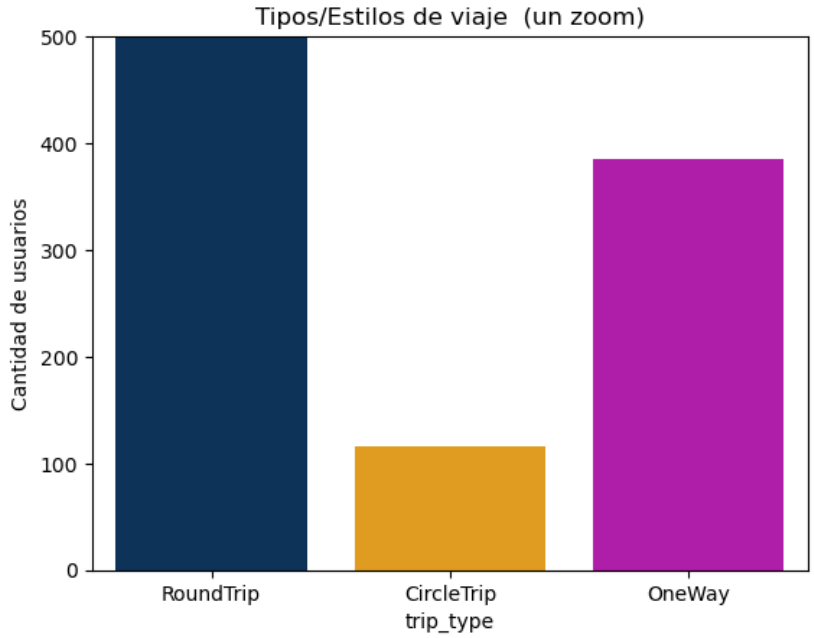
- No encontramos datos faltantes
- No encontramos filas repetidas
- Respecto a la columna numerica, no encontramos valores fuera de lugar, siempr en el rango 1 - 10

4. Exploracion






Tipos/Estilos de viaje
50000
40000
30000
20000
10000
0
Cantidad de usuarios
RoundTrip
CircleTrip
OneWay
trip_type


Tipos/Estilos de viaje (un zoom)
500
400
300
200
100
0
Cantidad de usuarios
RoundTrip
CircleTrip
OneWay
trip_type

De los graficos y con ayuda de pandas python obtenemos:

- Canal de venta:
  - Internet: 43842 usuarios ->89.10 %
  - Mobile: 5361 usuarios ->10.90 %
- Dia de reserva:
  - Lunes: 7974 usuarios ->16.206329 %
  - Martes: 7545 usuarios ->15.334431 %
  - Miercoles: 7548 usuarios ->15.340528 %
  - Jueves: 7310 usuarios ->14.856818 %
  - Viernes: 6674 usuarios ->13.564214 %
  - Sabado: 5722 usuarios ->11.629372 %
  - Domingo: 6430 usuarios ->13.068309 %
- Estilo de viaje:
  - Ida y vuelta: 48702 usuarios ->98.981769 %
  - Ida: 385 usuarios ->0.782471
  - Por tramos: 116 usuarios ->0.235758 %

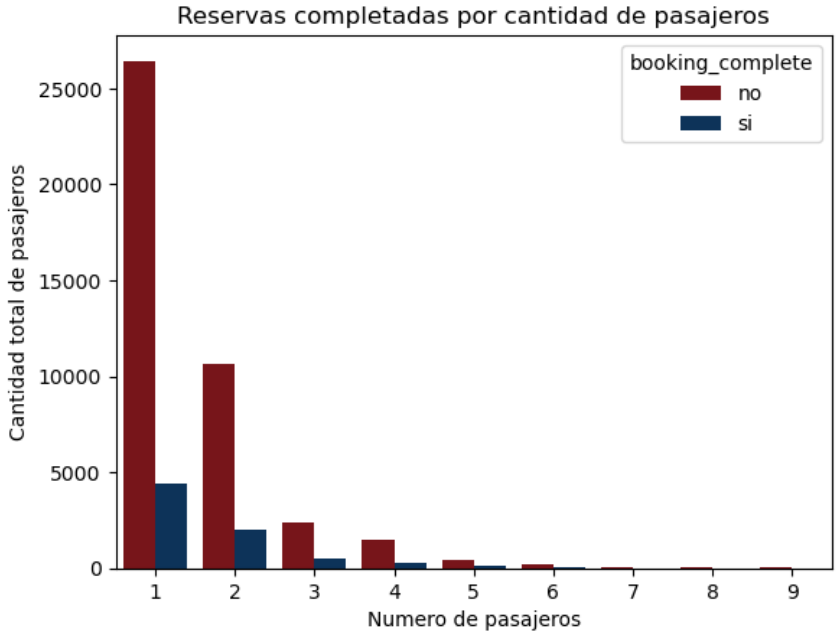
Primeras conclusiones:

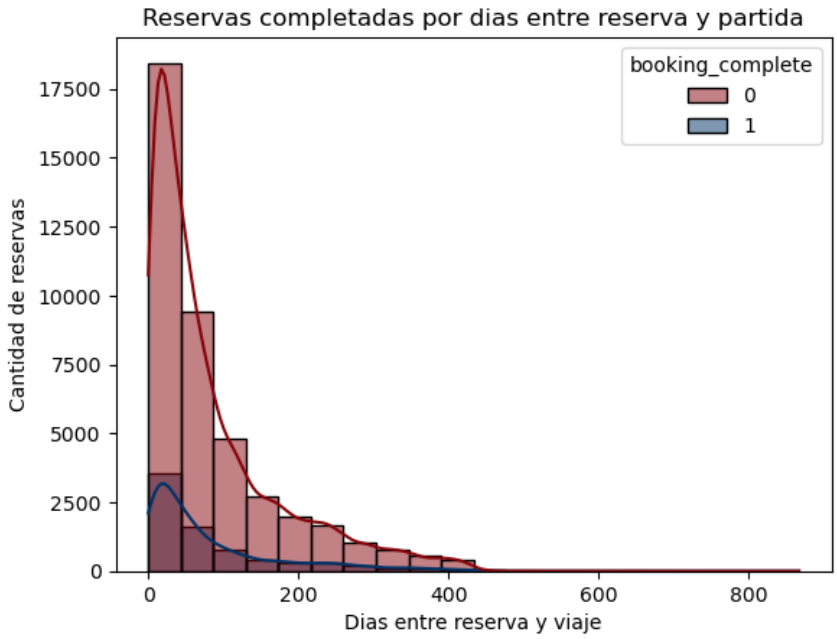
1. Dominio de Internet como canal de venta.
2. Viajes de ida y vuelta es la opción favorita por una amplia mayoría en las reservas. Esta tendencia indica que los clientes buscan viajes completos, tanto de ida como de regreso.
3. Equilibrio en la elección de días. Se observa un equilibrio en los días seleccionados para las reservas, con el lunes como el día con mayor número de reservas. Esto sugiere que no hay un día de la semana que sea claramente preferido por los clientes para realizar sus reservas.

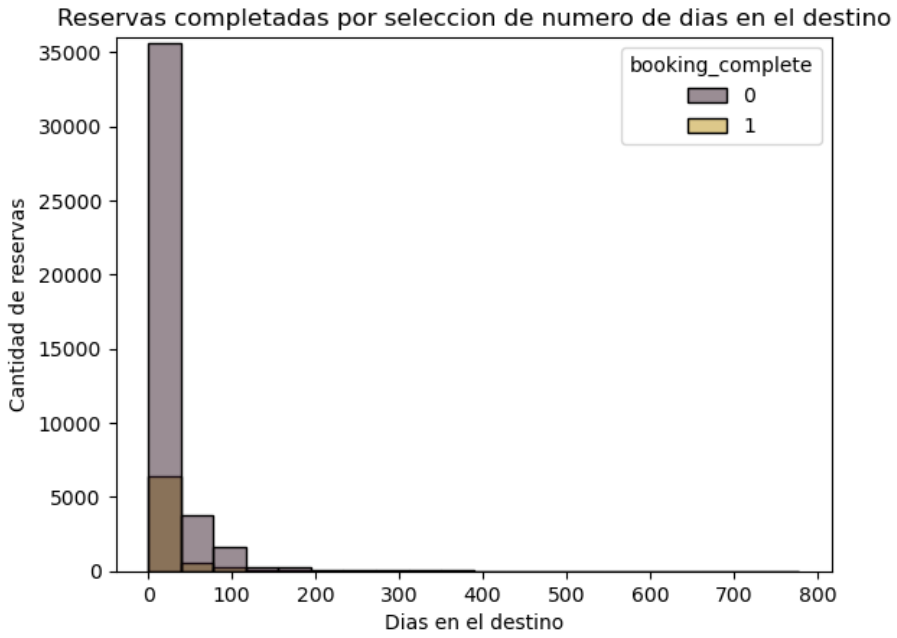
5. Pregunta de Interes

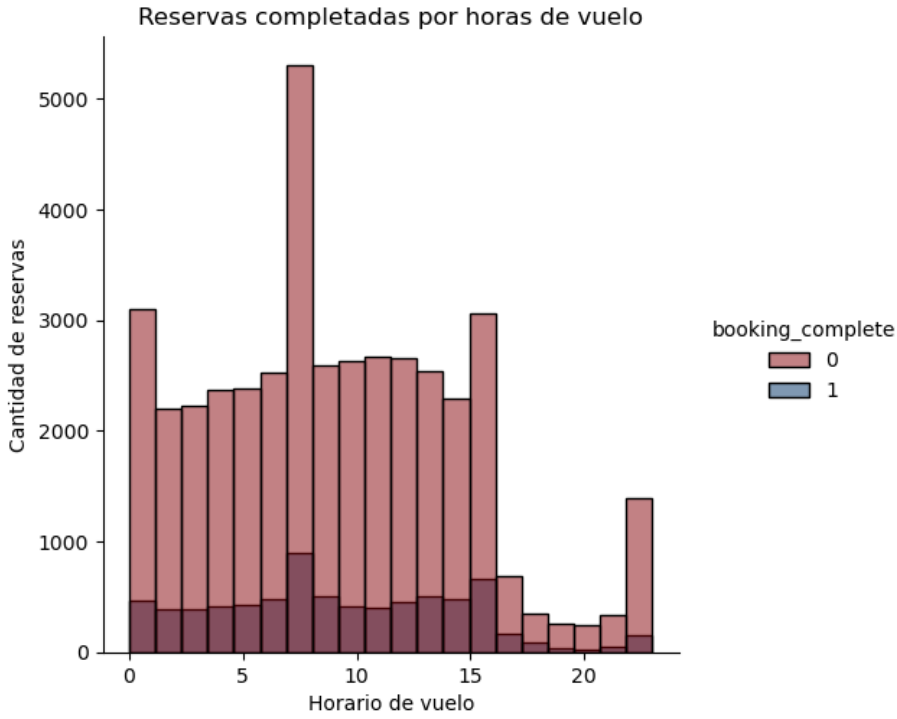
Analizaremos la posibilidad de concretar una reserva y las variables que influyen en la misma

6. Analisis


Reservas completadas por cantidad de pasajeros
booking_complete
no
si
Cantidad total de pasajeros
25000
20000
15000
10000
5000
0
1
2
3
4
5
6
7
8
9
Numero de pasajeros


Reservas completadas por dias entre reserva y partida
booking_complete
0
1
Cantidad de reservas
17500
15000
12500
10000
7500
5000
2500
0
0
200
400
600
800
Dias entre reserva y viaje



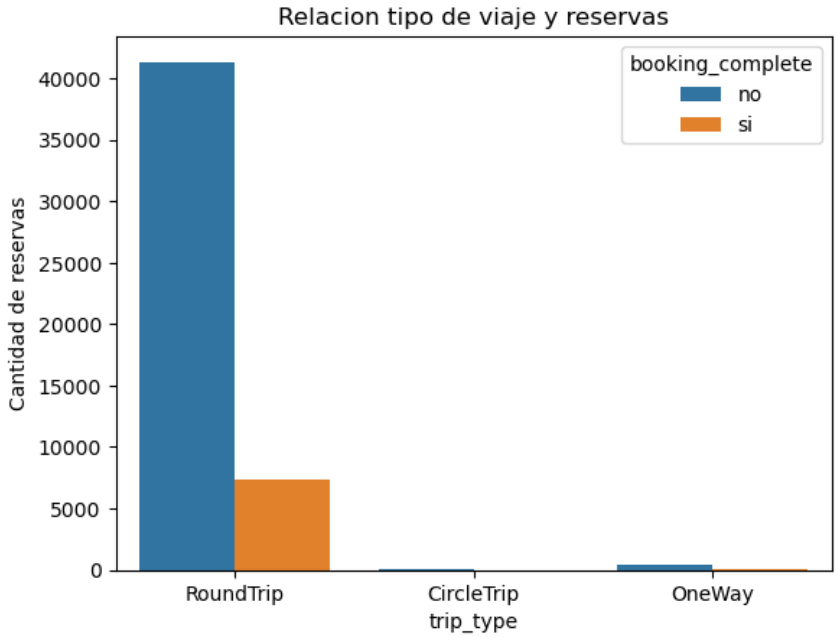


Análisis de las reservas:

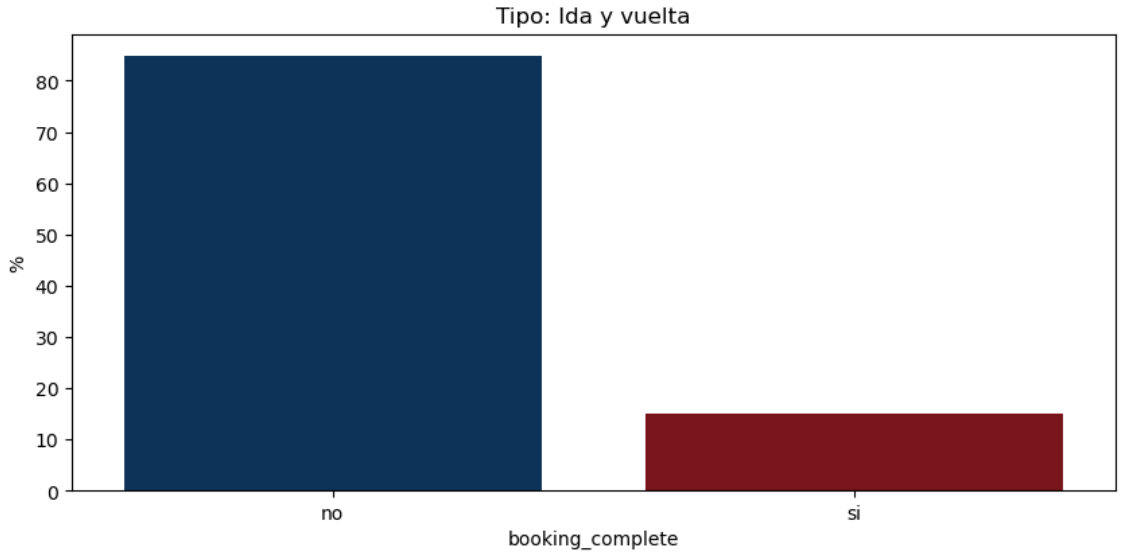
- **Proporción:** Las reservas completadas y no completadas son distribuciones equivalentes.
- **Tamaño del grupo:** Las reservas completadas son más frecuentes para grupos pequeños (no más de 3 pasajeros).
- **Tiempo de espera:** Las reservas completadas se realizan con una espera entre 0 y 100 días antes del viaje.

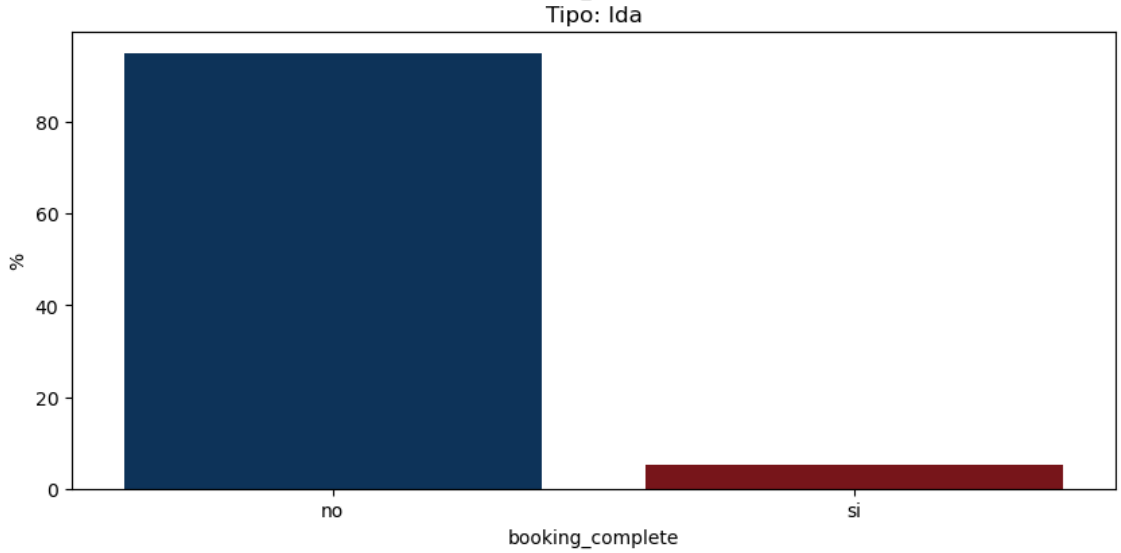
- **Duración del viaje:** Las reservas completadas se realizan para viajes de no más de 45 días en el destino.
- **Horario de partida:** Las reservas completadas son más frecuentes para partidas a las 7:00 a.m.

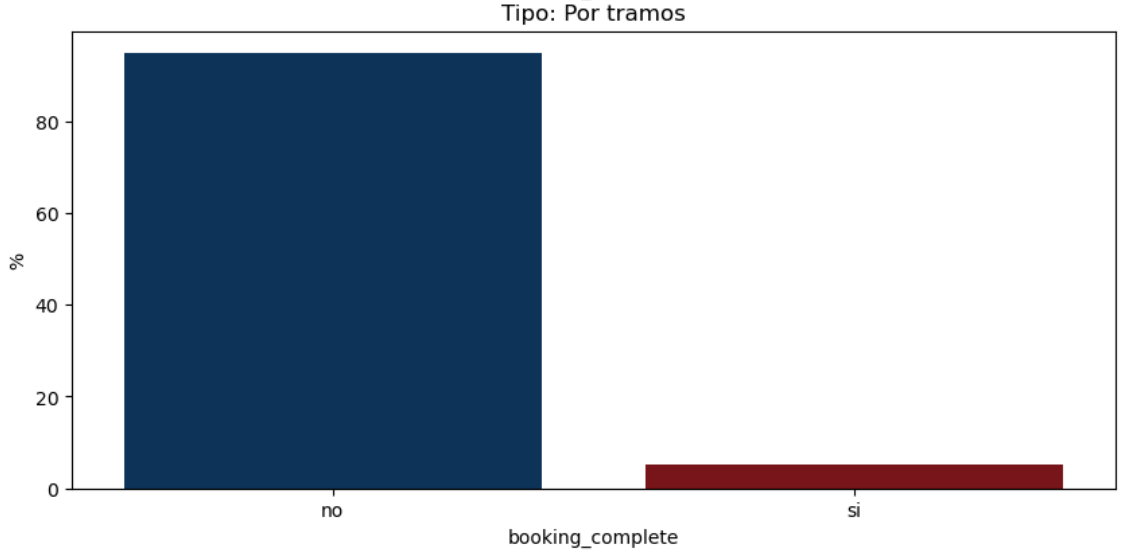
Es importante destacar que estos son solo algunos puntos clave del análisis. Para obtener una comprensión más completa, aumentaremos el analisis poniendolo en funcion de otras variables.



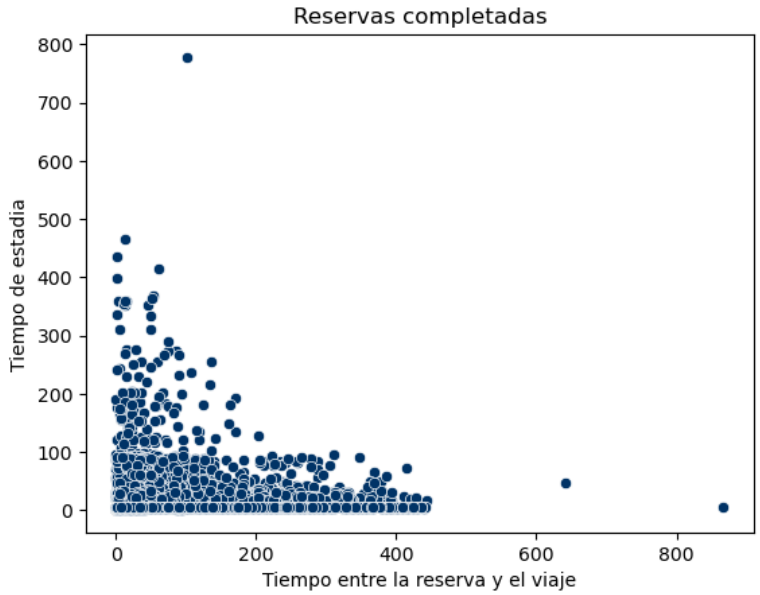
Con anterioridad descubrimos que en la cantidad de reservas por tipo de viaje era ampliamente superdora la opcion de RoundTrip, con el grafico mostrado, discriminamos entre las completadas y no completadas. Es lògico que la cantidad de reservas completadas va a ser desproporcionada por la cantidad de datos que ingresa a cada set, pero, què ocurre en valores porcentuales ?

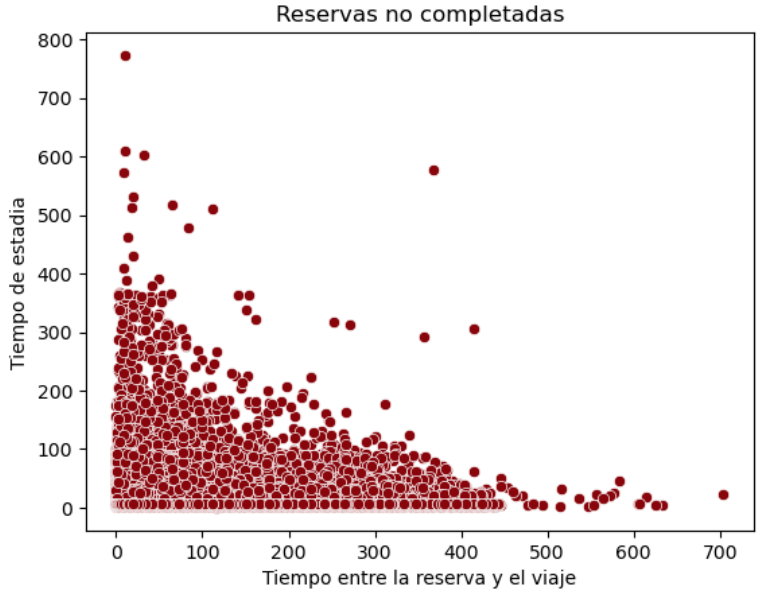

Tipo: Ida y vuelta
%
0
10
20
30
40
50
60
70
80
no
si
booking_complete


Tipo: Ida
%
0
20
40
60
80
no
si
booking_complete


Tipo: Por tramos
%
0
20
40
60
80
no
si
booking_complete

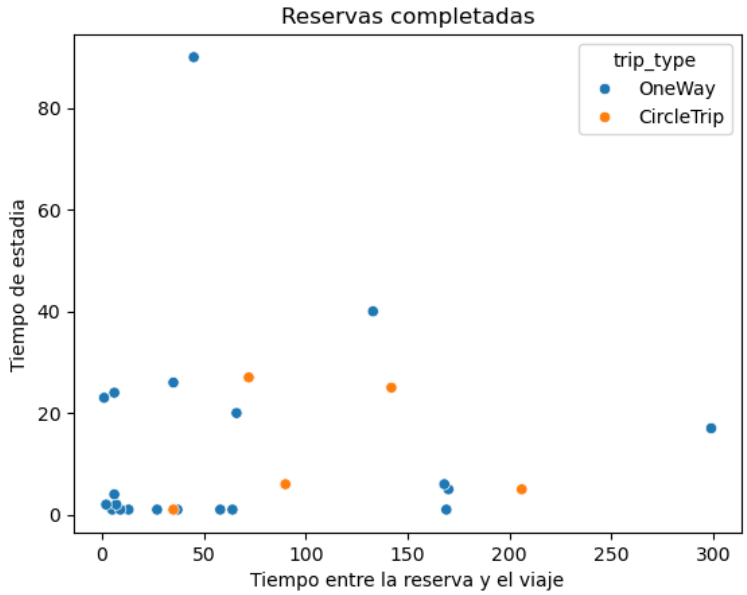
Com se puede ver, los porcentajes se mantienen proporcionales, mostrando aùn roundtrip el tipo de viaje mas solicitado. Pasamos a ver que ipo de correlacion hay entre el tiempo de estadia en el destino y el tiempo de espera entre el viaje y la reserva:



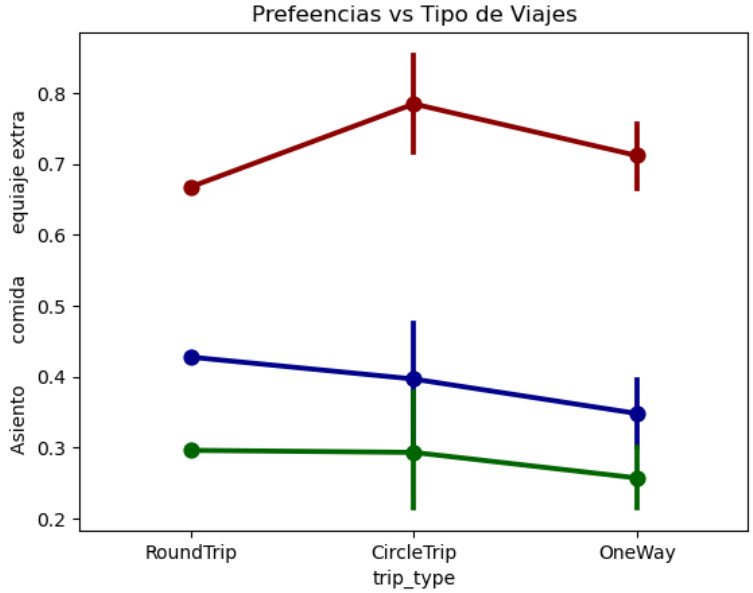


La distribucion se hace mas masiva cuanto mas al origen nos acercamos. Cuanto menos es el tiempo de viaje y reserva, el tiempo de estadia tiee una distribucion mas equitativa. Pero si nos alejamos (un zoom out) veremos que a medida que disminuimos mas el tiempo entre reserva y viaje, mayor es el tiempo de estadis, lo que sugiere que hay una correlacion negativa entre las variables. El gráfico de estilos de viaje nos reveló que el viaje RoundTrip es la opción predilecta. Pero, ¿a qué se debe esta preferencia? Descubriendo las claves del roundtrip: Estancias cortas: La mayoría de las reservas se realizan para estancias medianamente cortas, inferiores a 100 días. El tiempo de reserva y el tiempo de viaje es un factor determinante.

A continuacion, mostramos la distribucion de reservas completadas para los tipos OneWay y CircleTrip, demostrando que no hay correlacion entre estas, por lo tanto la relacion mencionada anteriormente es exclusivamente de RoundTrip

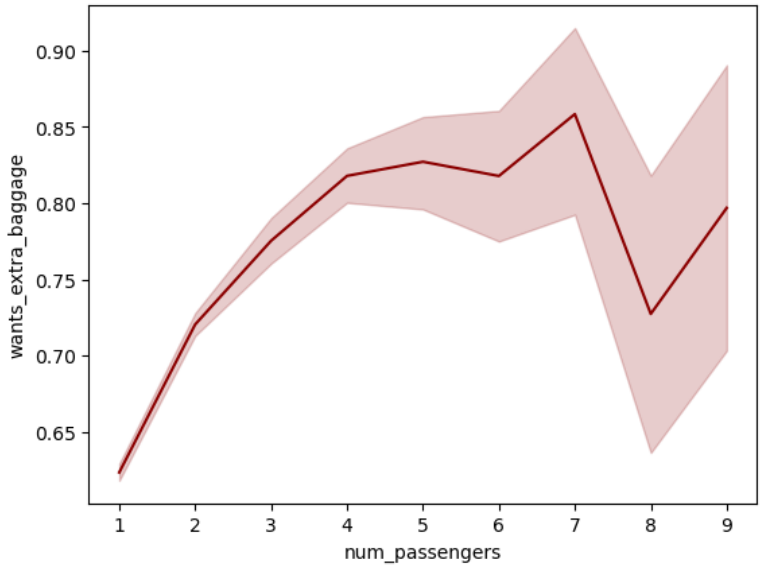


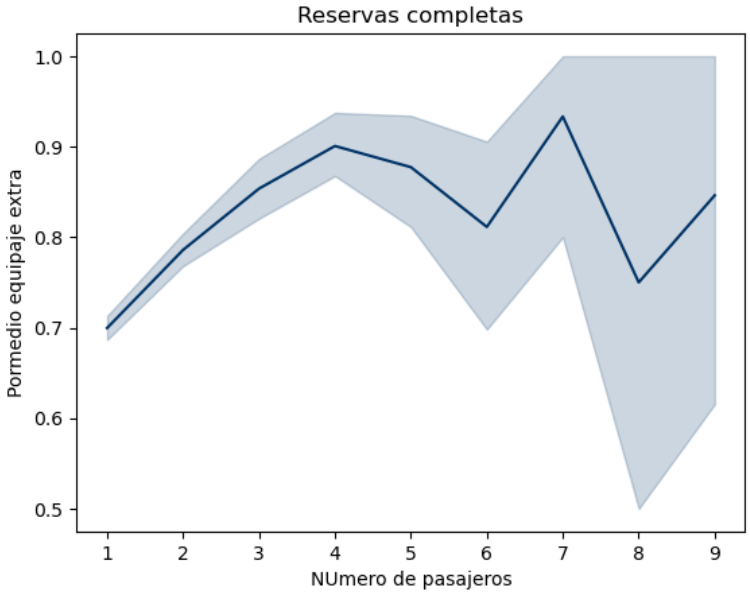
Preferencias: No se observa una relación significativa entre la preferencia del sitio y el tipo de viaje, por lo que podemos decir que la preferencia del sitio no parece ser un factor determinante en la elección del tipo de viaje ya que en promedio no varia cuando miramos los trres tipos de viaje disponibles Comida preferida se solicita con menos frecuencia en viajes de OneWay, respecto a los otros dos tipos, manteniendo el promedio de solicitudes inferior al 50 % aunque siempre tiene mayor solicitud que el asiento preferido. En cuanto al el equipaje extra observamos un notable aumento en la solicitud en viajes por tramos (CircleTrip). El promedio de solicitudes de equipaje extra supera el 60 % en todos los tipos de viaje.



Es interesante analizar si disminucion del equipaje extra se debe a que el numero

de solicitante de reserva es mayor:





7. Apendice