TRABAJO PRÁCTICO

Grupo Data Minds: Agustina Fernández, Irina Di Fonzo y Barbara Kildal.

1. Selección del Dataset asignado: revisá según tu equipo que dataset tenes que utilizar.

Nuestro Dataset está relacionado con los datos almacenados por un hotel y su sistema de reservas.

2. Definición del Problema: deberán definir un problema específico de negocio o investigación que quieran resolver. Esta definición debe incluir una clara formulación de qué quieren predecir o clasificar y por qué es importante o relevante.

Nuestro problema de negocio hace referencia a la cancelación de las reservas, o inasistencias por parte de los huéspedes. Queremos intentar comprender cuál puede ser el motivo de estas cancelaciones, y si hay alguna manera de predecir si un futuro huésped cancelará su reserva. Esta predicción es relevante porque permite minimizar la pérdida de ingresos, mejorar la planificación de recursos y optimizar la ocupación del alojamiento.

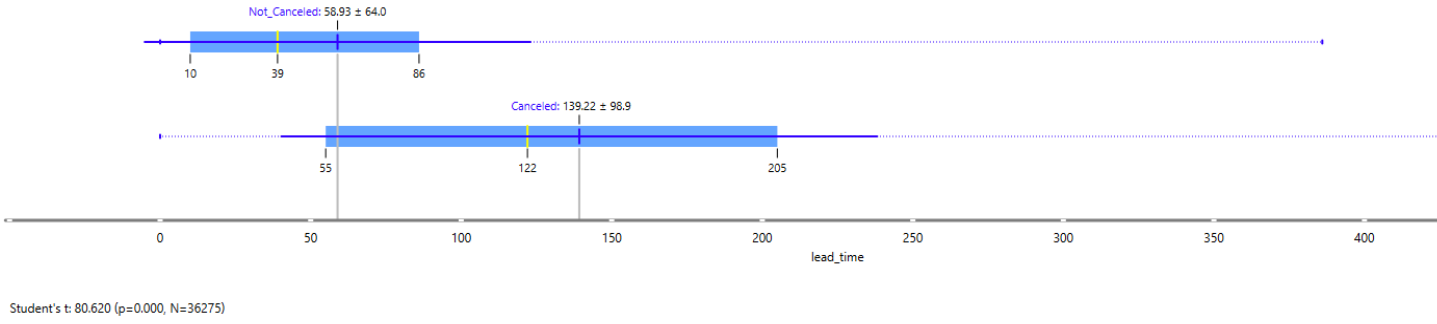
3. Análisis Exploratorio de Datos:

1. **Realicen un análisis exploratorio inicial para familiarizarse con los datos.**
2. **Planteen hipótesis sobre qué variables creen que son más relevantes para predecir el resultado de interés.**
3. **Utilicen gráficos y estadísticas para apoyar estas hipótesis y documenten sus hallazgos.**

1. Nuestra base de datos contiene variables como tipo de habitación, número de adultos o niños, cantidad de noches reservadas, plan de comidas elegido, si necesita estacionamiento, y el historial del huésped, referidas a sus reservas y cancelaciones pasadas, entre otras. Alojan toda la información necesaria para poder generar una reserva en el sistema y mantener un seguimiento de aquellos huéspedes que vuelven a reservar en el hotel.
2. Consideramos que para poder desarrollar un modelo que prediga futuras cancelaciones, las siguientes variables son relevantes:
    - lead_time: esta variable hace referencia a los días entre la reserva y el día de llegada del huésped al hotel. Consideramos relevante saber qué tan cerca de la fecha de la llegada estamos, porque a menor tiempo, menos posibilidad de cancelación.
    - arrival_month: tomamos el mes de llegada como relevante ya que puede existir el caso en el que futuros huéspedes quieran asegurarse su lugar en el hotel en una fecha de temporada alta, y más adelante planifican mejor su viaje, evaluando otras alternativas, y de ser necesario, cancelando la reserva previamente hecha, sin cargo.

- repeated_guest: consideramos que si se trata de un huésped que ya haya visitado el hotel antes, hay menos posibilidad de cancelación, porque ya conoce el servicio y volverá a escoger al hotel como destino de alojamiento.
- no_of_previous_cancellations: el número de reservas canceladas previamente por un mismo huésped nos da información acerca de las tendencias de los huéspedes, y qué tan frecuente es la cancelación de reservas, permitiéndonos predecir un comportamiento futuro.
- no_of_previous_bookings_not_canceled: podemos utilizar esta variable para relacionarla con el número de cancelaciones, permite a calcular la diferencia de huéspedes que cancelan ocasionalmente de aquellos que rara vez cancelan.
- no_of_special_requests: esta variable puede representar el nivel de planificación que el huésped realizó con la reserva, y pensamos que cuantas más especificaciones, menor probabilidad de cancelación o no show.
- booking_status: nos ayuda a relacionar variables, y saber el estado actual de las reservas.
  No tenemos en consideración variables como el tipo de habitación o el precio promedio por habitación, ya que se menciona que las reservas son de cancelación gratuita o tiene un precio bajo que beneficia a los huéspedes.

Gráfico 1: Relación entre lead time y booking status utilizando un box plot.



Pudimos ver que las reservas canceladas tienen tiempos de anticipación mucho mayores que las no canceladas.
El Q3 de Not_Canceled (86 días) está por debajo del Q1 de Canceled (122 días). Eso indica que a medida que crece lead_time, aumenta la probabilidad de cancelación.

Gráfico 2: Relación entre arrival month y booking status para analizar la estacionalidad, utilizando un pivot table.

booking_status

arrival_month	Count	Canceled	Not_Canceled	Total
	1	24.0	990.0	**1014.0**
	2	430.0	1274.0	**1704.0**
	3	700.0	1658.0	**2358.0**
	4	995.0	1741.0	**2736.0**
	5	948.0	1650.0	**2598.0**
	6	1291.0	1912.0	**3203.0**
	7	1314.0	1606.0	**2920.0**
	8	1488.0	2325.0	**3813.0**
	9	1538.0	3073.0	**4611.0**
	10	1880.0	3437.0	**5317.0**
	11	875.0	2105.0	**2980.0**
	12	402.0	2619.0	**3021.0**
	Total	**11885.0**	**24390.0**	**36275.0**

Analizamos la estacionalidad calculando la tasa de cancelación, haciendo candeled / total, para cada mes.

Count	Canceled	Not_Canceled	Total	Tasa de cancelación
1	24	990	1014	2,37%
2	430	1274	1704	25,23%
3	700	1658	2358	29,69%
4	995	1741	2736	36,37%
5	948	1650	2598	36,49%
6	1291	1912	3203	40,31%
7	1314	1606	2920	45,00%
8	1488	2325	3813	39,02%
9	1538	3073	4611	33,36%
10	1880	3437	5317	35,36%
11	875	2105	2980	29,36%
12	402	2619	3021	13,31%
Total	11885	24390	36275	

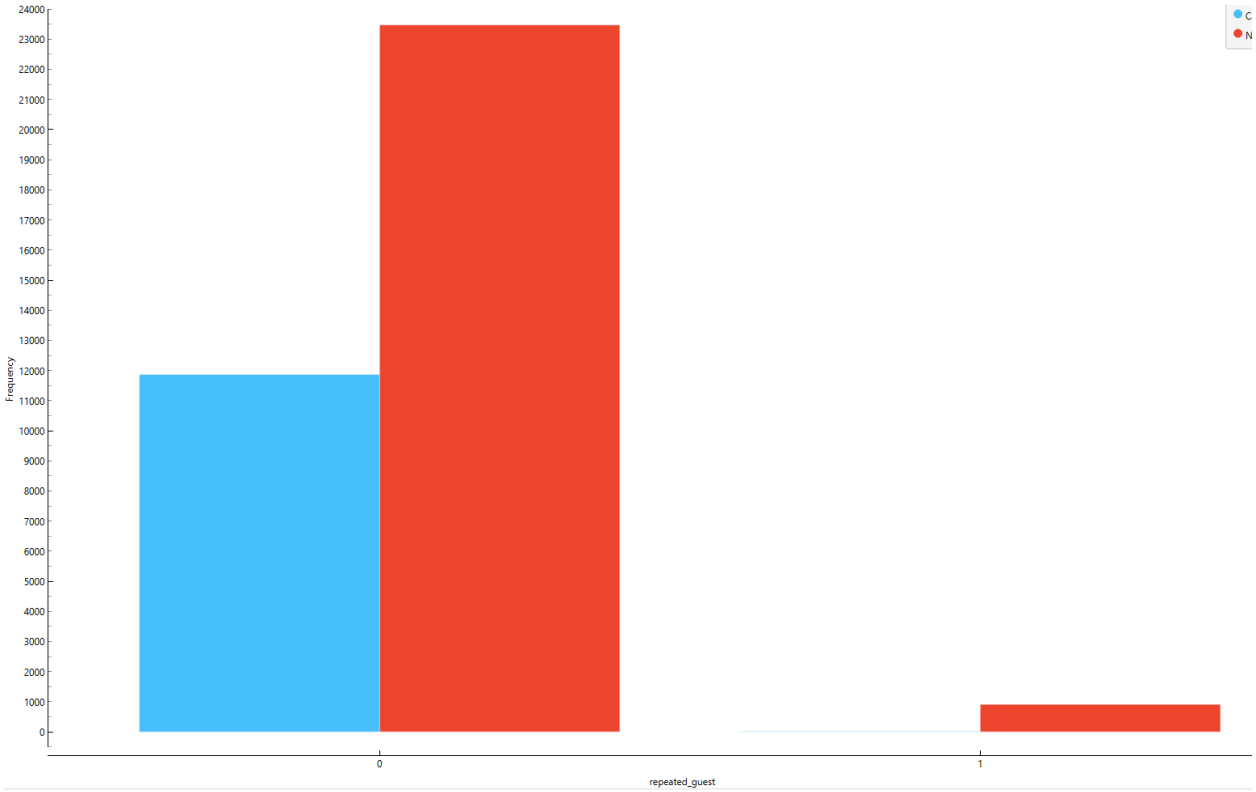
De junio a agosto, las tasas de cancelación están por encima del promedio entre 40% y 45%, con julio siendo el más alto. Esto indica que se tratan de meses de alta demanda o vacaciones, por ende las reservas están hechas con mucha anticipación para “asegurar lugar” y luego

replanificadas o reemplazadas (coherente con el gráfico anterior que muestra a mayor lead time, más cancelaciones).

Por otro lado, enero y diciembre están muy por debajo del promedio.

Existe estacionalidad en la cancelación: pico en mitad de año, mínimos en enero/diciembre.

Gráfico 3: Relación entre los huéspedes repetidos y booking status, para analizar si aquellos que vuelven al hotel son menos propensos a cancelar.

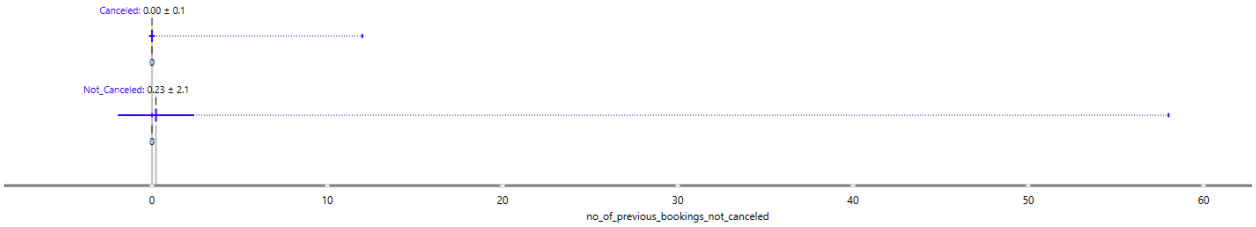


La inmensa mayoría de las reservas son de huéspedes NO reincidentes (repeated_guest = 0). En ese grupo aparecen ambos colores: muchas “Not_Canceled” y también un volumen importante de canceladas.

Entre los reincidentes (repeated_guest = 1) casi todo es “Not_Canceled” y la barra azul (Canceled) es prácticamente inexistente.

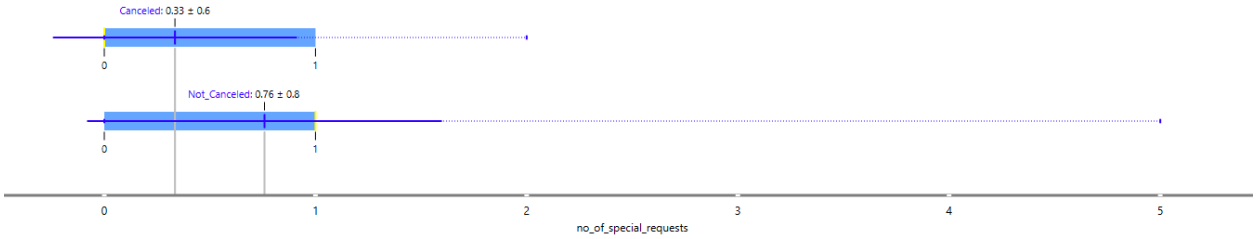
Esto nos hace entender que ser huésped reincidente se asocia con mucha menor probabilidad de cancelar, es decir, entre los reincidentes, la cancelación es muy rara (la distribución muestra una barra casi nula de “Canceled”).
Podemos concluir que la experiencia previa con el hotel parece proteger contra la cancelación, por lo que repeated_guest es un predictor negativo de “Canceled”.

Gráfico 4: Box plot que relaciona el número de reservas sin cancelaciones previas y el booking status.



Entendemos que tener al menos 1 reserva previa no cancelada es mucho más común entre las reservas no canceladas actuales.
Esto indica que, cuando existe historial de cumplimiento, la probabilidad de no cancelar aumenta. Por ende, contar con historial positivo previo se asocia con menor probabilidad de cancelación.

Gráfico 5: box plot que relaciona el número de pedidos especiales con el booking status.



Este gráfico nos indica que a mayor cantidad de pedidos especiales, menor cancelación. Las reservas no canceladas tienden a tener más special_requests que las canceladas. Esto puede indicar que los pedidos especiales indican mayor planeación del viaje, lo que disminuye la probabilidad de cancelar.