

HOTEL RESERVATIONS EDA

Grupo 5

Manuel García Plaza
José Miguel Ramírez Muñoz

Índice

1. Vista general

- 1.1. Resumen ejecutivo
- 1.2. Objetivos
- 1.3. Qué datos hay

2. Precios

- 2.1. Tipo de habitación
- 2.2. Fecha
- 2.3. Otros

3. Familias

- 3.1. Con/sin niños

4. Cancelaciones

- 4.1. Historial
- 4.2. Precio
- 4.3. Forma de reserva

5. Conclusiones

6. Next Steps

7. Backup

Resumen ejecutivo

DATOS

Trabajamos con 36.275 individuos y 18 variables, de las cuales 12 son numéricas y 6 son categóricas. Las más relevantes son:

- Precio medio por habitación.
- Edad.
- Cancelación de la reserva.

EDA

Visualizamos estas variables de interés en histogramas y gráficas de barras para facilitar su interpretación. Comprobamos si pudiera existir una posible relación entre ellas antes de realizar algún contraste de hipótesis.

CONCLUSIONES

Habiendo trabajado los datos, realizamos nuestras conclusiones y especificamos los “*next steps*”, es decir, los posibles pasos posteriores sobre cómo actuar y posibles recomendaciones de cara a las demandas del mercado.

Objetivos:

Principalmente identificar patrones de las cancelaciones para poder minimizarlas o compensarlas; además, observar comportamiento del negocio en líneas generales.

Para ello:

- Proponer variables de interés.
- Intentar encontrar ciertas relaciones entre variables o entre grupos de individuos según alguna de estas variables.
- Intentar encontrar outliers, individuos con valores que no tengan sentido.
- Depurar el conjunto de datos.

¿Qué datos tenemos?

Fuente:

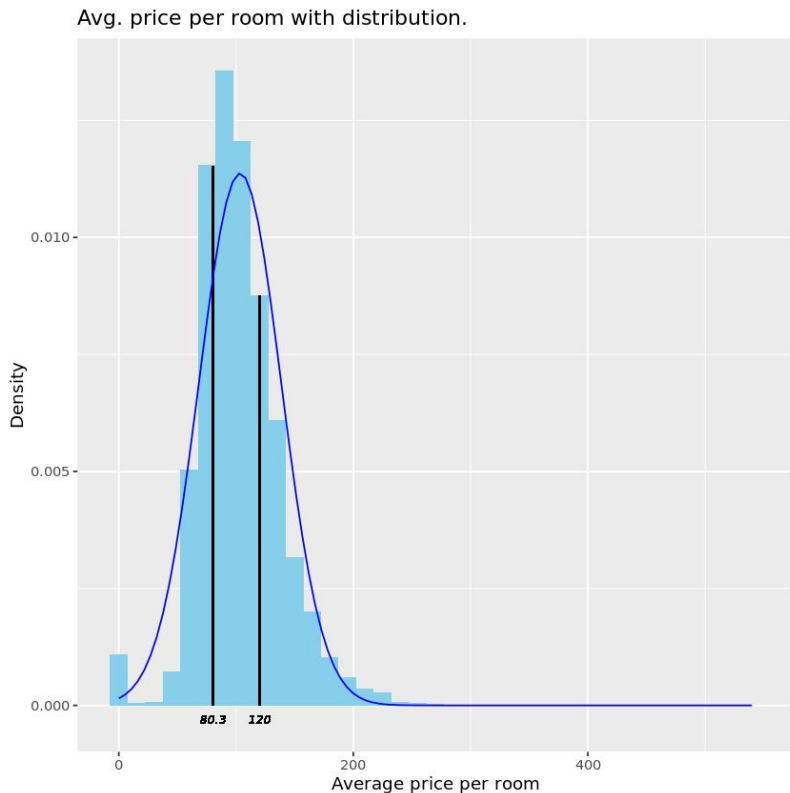
<https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset>

Cada fila corresponde a una reserva de un hotel, es decir, un cliente. De esta se recogen:

- Número de adultos y niños.
- Número de noches de diario y fin de semana.
- Pensión (desayuno/media/completa).
- Parking (Sí/No).
- Tipo habitación (7 tipos).
- Fecha de llegada (día, mes, año).
- Antelación de la reserva.
- Forma de reservar (5 formas).
- Número de cancelaciones/no cancelaciones previas.
- Precio medio por habitación.
- Número de requisitos especiales.
- Estado actual reserva.

No hay ninguna celda vacía ni con valor inválido. A priori no hay relación lineal entre ninguna variable (todas las correlaciones son menores que 0,7).

Distribución del precio medio por habitación:



Siguen una distribución normal:

- $\mu = 103.42\text{€}$
- $\sigma = 35.09\text{€}$

La curva superpuesta es la densidad de una normal con estos parámetros.

Las líneas verticales delimitan los tres sectores de precios definidos:

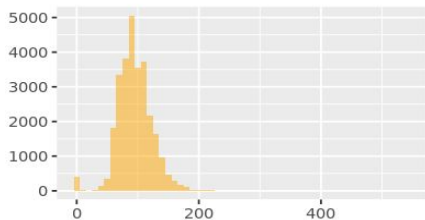
1. **Sector 1:** 0€ - 80.3€ (25%)
2. **Sector 2:** 80.3€ - 120€ (50%)
3. **Sector 3:** >120€ (25%)

Alerta: cantidad sorprendentemente alta de **valores cercanos a 0**.

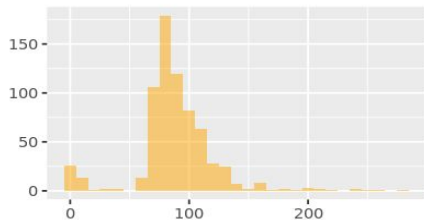
Precios

Hay semejanzas según el tipo de habitación por pares.

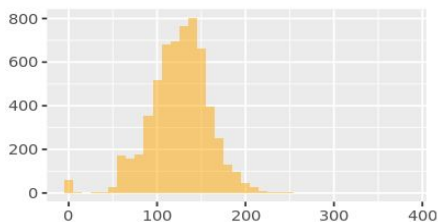
Room type 1



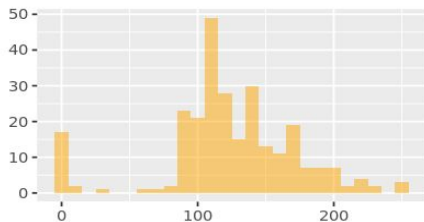
Room type 2



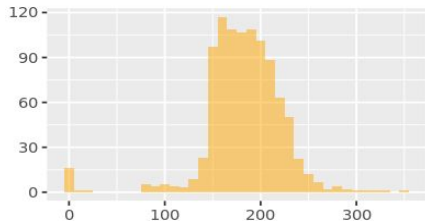
Room type 4



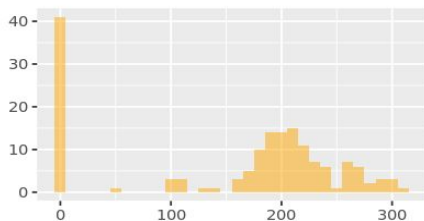
Room type 5



Room type 6



Room type 7



No incluimos tipo 3, solo hay 7 observaciones. Del resto tenemos más de 150.

Todas parecen normales aunque con diferentes medias y dispersiones.

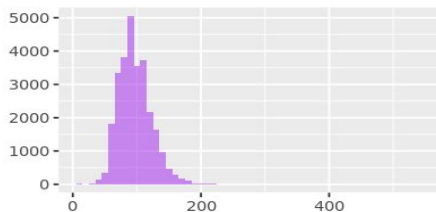
La 1 y la 2 son las más **baratas**, la 4 y la 5 son de precio **medio** y la 6 y la 7 las más **caras**.

En todas ellas se repiten los casos de precios próximos a cero, sobre todo en tipo 7.

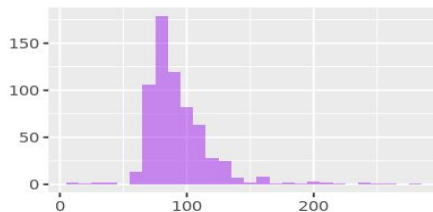
Precios

Eliminando los precios menores que 10€:

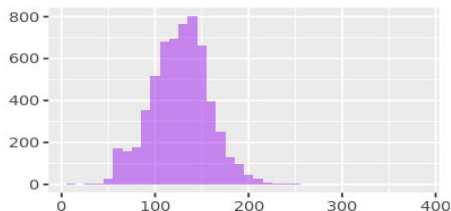
Room type 1



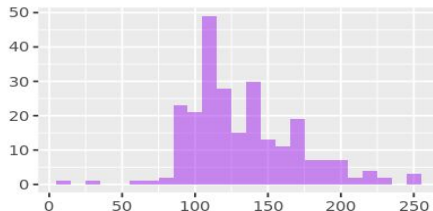
Room type 2



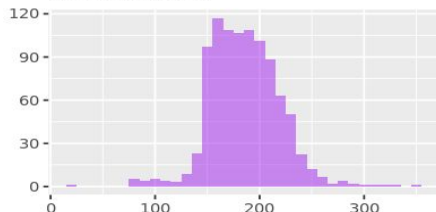
Room type 4



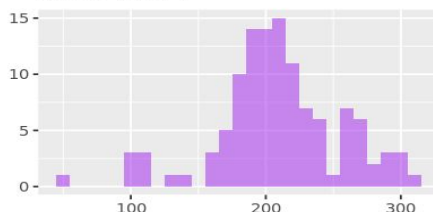
Room type 5



Room type 6



Room type 7



Sin estos valores raros (posibles errores en la toma de datos) las distribuciones siguen siendo prácticamente normales con medias similares, pero **¿realmente influyen?**

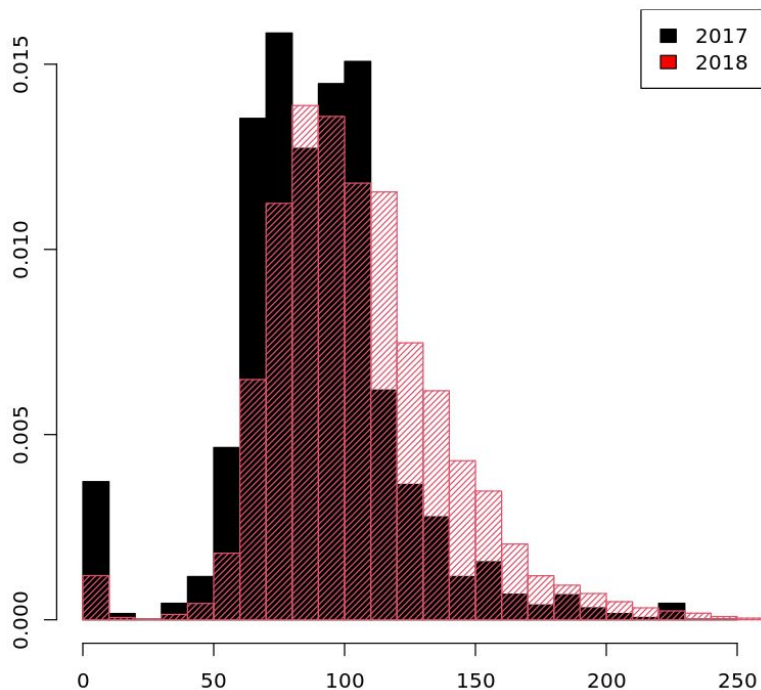
Lo vemos a continuación:

Precios

Estos valores influyen: incremento del 35% en el tipo 7.

Tipo Habitación	Media original	Media sin ceros	Diferencia medias	Porcentaje de diferencia
1	95.92	97.37	1.46	1.52
2	87.85	92.70	4.85	5.52
4	125.29	126.56	1.27	1.01
5	123.73	132.73	8.99	7.27
6	182.21	185.47	3.26	1.79
7	155.20	209.58	54.38	35.04

Aumento de la media 16€ de 2017 a 2018.



Hay 6514 observaciones (18%) de 2017 y 29761 (82%) de 2018 (los únicos años de los que tenemos datos).

Ambos grupos tienen una distribución aproximadamente normal pero hay un incremento en la media de los precios medios:

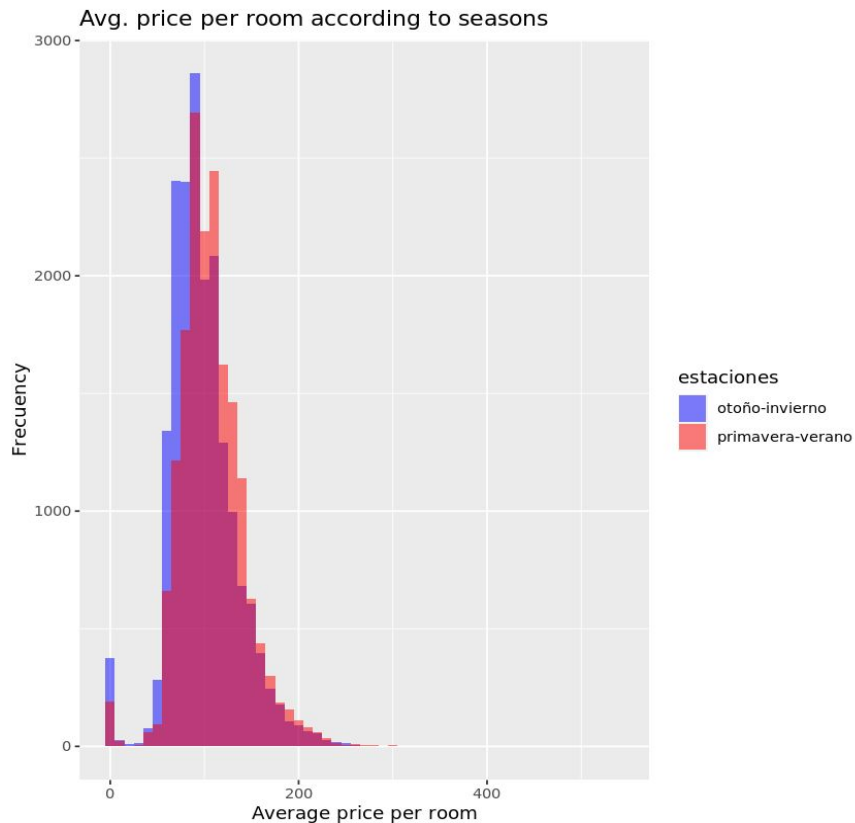
media 2017 = 90.03€

media 2018 = 106.36€

Enorme proporción de precios cercanos a 0 en 2017 (sin estos la media aumenta a 93.5€).

Precios

Entre estaciones los precios se mantienen.



Aquí se tienen de nuevo muestras distribuidas normalmente y con medias:

media o-i = 98.89€

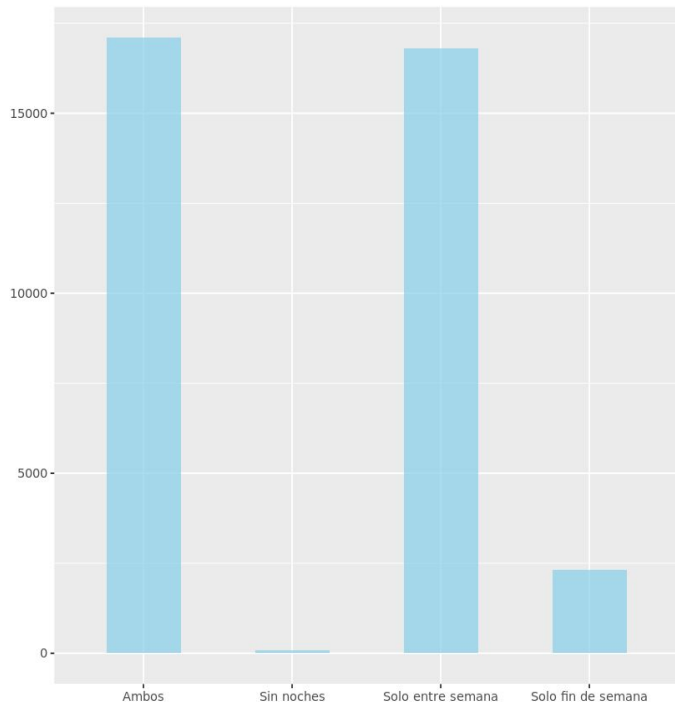
media p-v = 108.22€

Otoño-Invierno es donde ahora tenemos la mayor proporción de precios medios cercanos a cero (sin estos la media aumenta a 101.01€).

El 48.6% de las reservas son en Primavera-Verano; 51.4% son en Otoño-Invierno.

Precios

No hay diferencias notables entre los días de la semana.



La mayor parte de reservas son con días de los dos tipos o sin fines de semana. Hay poca proporción de únicamente fines de semana pero es suficiente cantidad.

Las **medias** de precios según el grupo son:

ambos: 103.67€

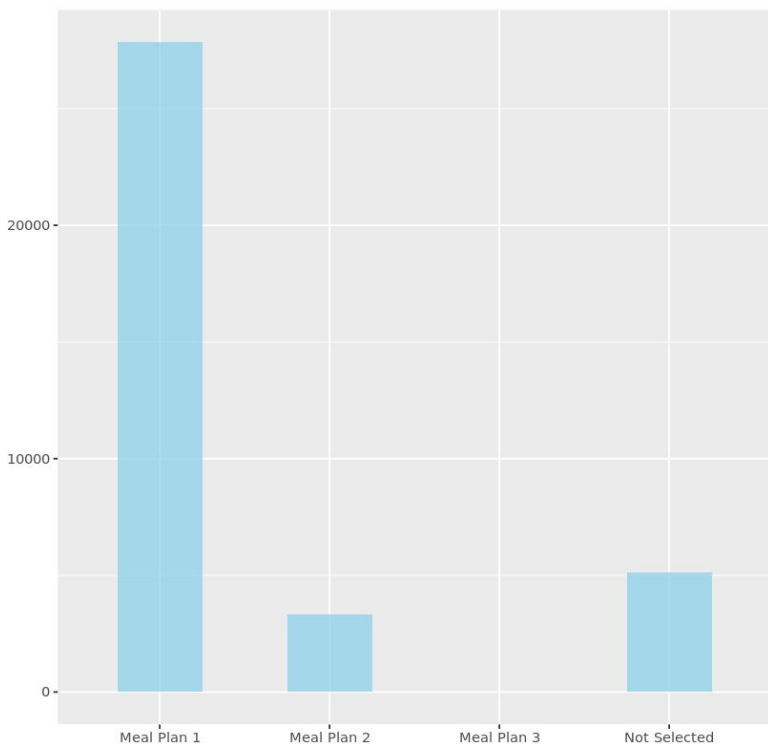
solo findes: 101.59€

solo diario: 103.90€

Sin diferencias notables entre estos grupos.

Precios

Los precios aumentan según el tipo de pensión.



Aquí tenemos la cantidad de reservas según el tipo de comida. Veamos las **medias** por grupos:

tipo 1: 103.59€

tipo 2: 115.31€

tipo 3: 41.2€

sin seleccionar: 94.91€

Alerta con tipo 3: debería ser la más cara (solo 5 observaciones y 4 de ellas son de precio medio 0).

Obviamente tipo 2 es más caro que el 1 y estos 2 más caros que sin plan.

Precios

Se gasta más en familia, pero al reservar con antelación no cambia nada.

	Con niños	Sin niños		Más de 14 días	14 días o menos
Media	144.29	100.40		103.93	101.74
Frecuencias	2559	33577		27890	8385

En cuanto a viajar con o sin niños parece haber **diferencia**; se gasta **más con niños**.

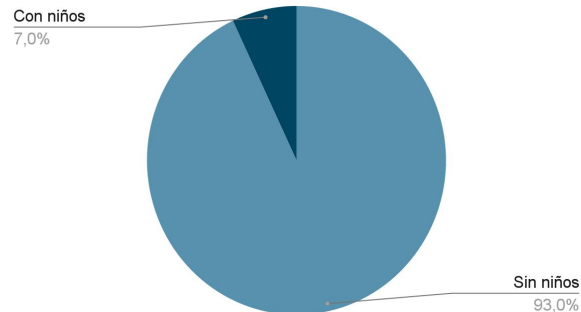
Respecto a la **antelación**, **no** parece haber **diferencia** relevante.

Familias

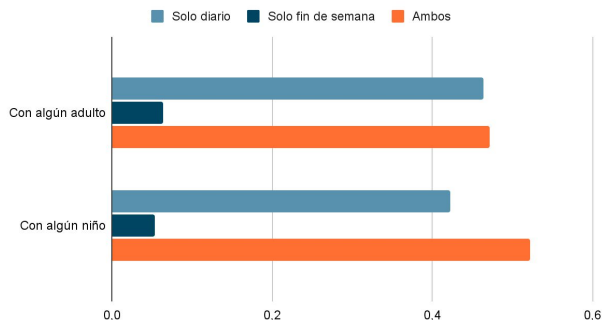
Preferencias de los clientes en función de ir con o sin niños:

En todas las reservas hay al menos una persona incluida (no hay reservas absurdas).

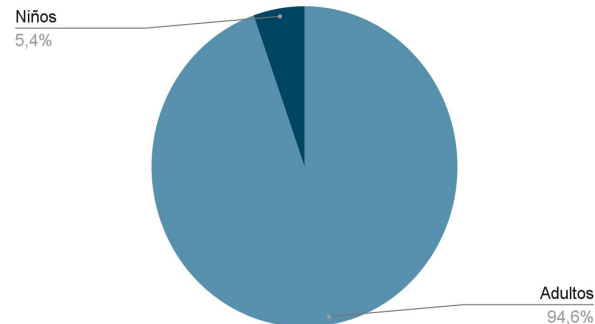
Reservas con adultos (99.6% del total de reservas)



Proporciones por días de la semana



Clientes totales

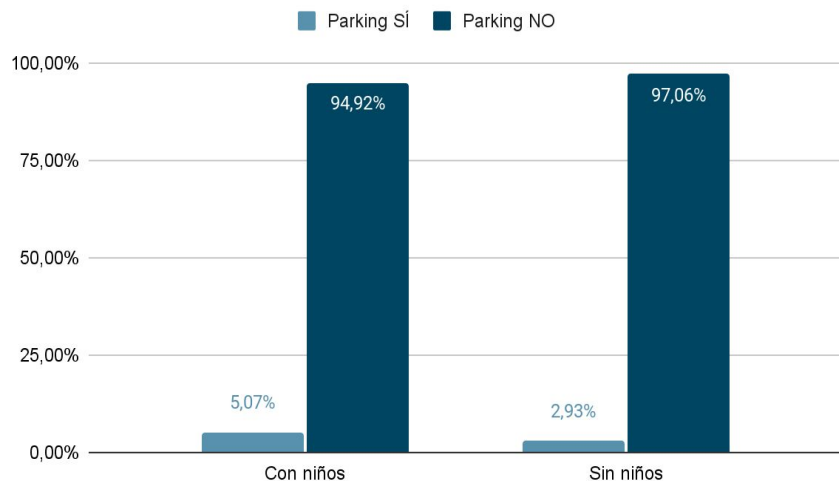


Es decir, **no** tenemos grandes **diferencias** separando los individuos por días de la semana. El público adulto es el mayoritario.

Familias

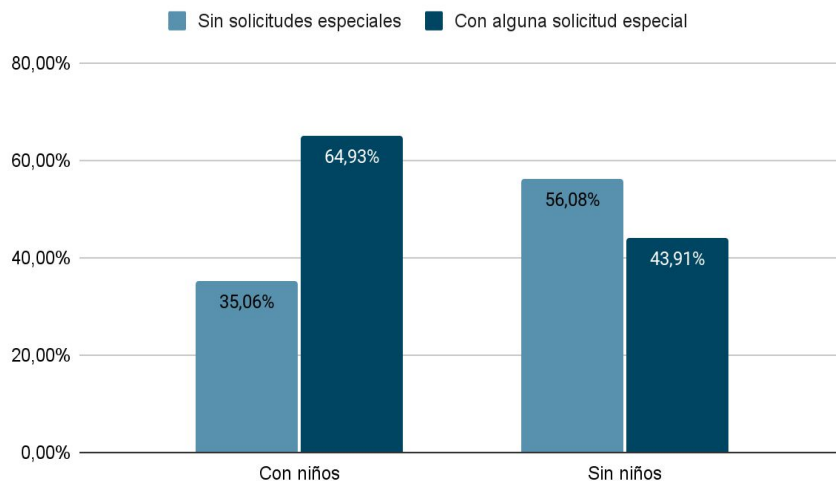
**El parking es muy poco solicitado en cualquier caso.
Más del 50% de las familias hacen solicitudes especiales.**

Proporciones (parking)



No hay diferencias entre aquellos que van **con niños** o **sin niños** a la hora de solicitar parking.

Proporciones (solicitudes especiales)

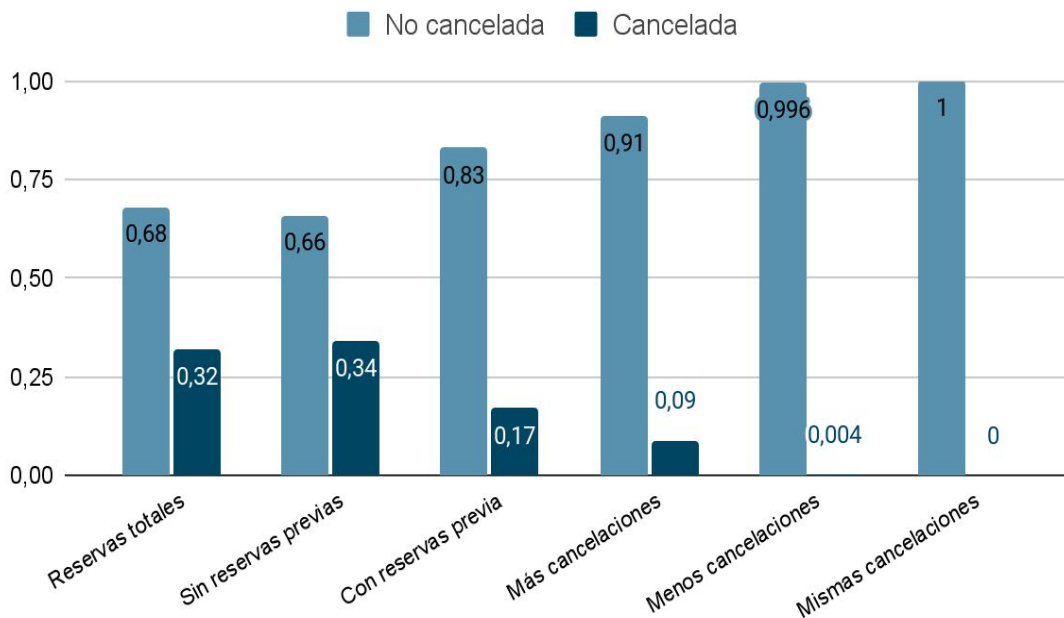


Sin embargo, observamos que los adultos que viajan **con niños** realizan un **65%** de solicitudes especiales frente a los otros

Cancelaciones

No se observan tendencias dispares según las reservas pasadas.

Proporción del estado actual de las cancelaciones



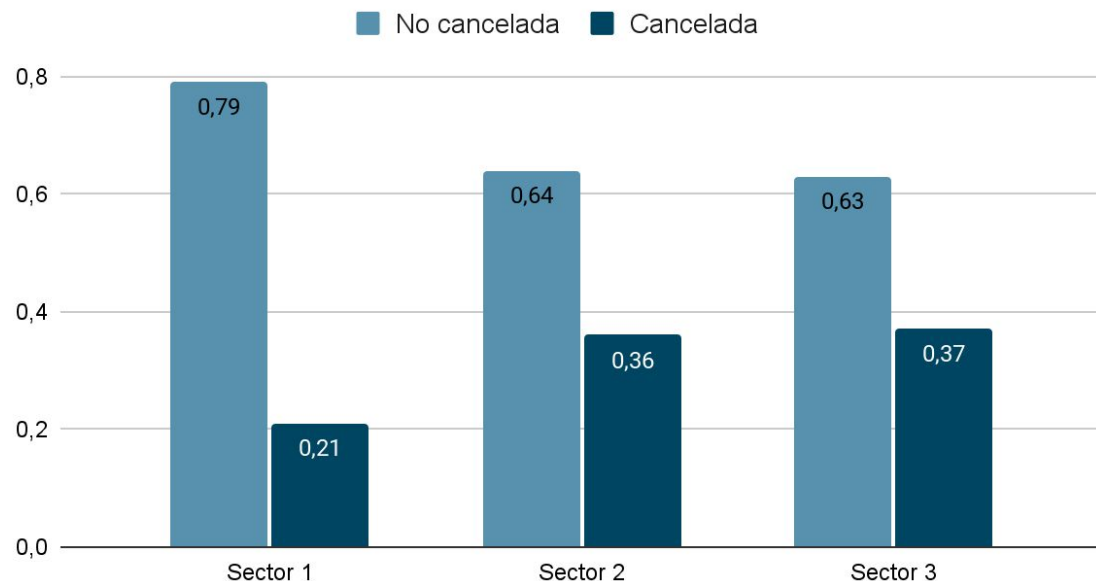
La masa **total** y la gente **nueva** se comporta prácticamente **igual**. Los clientes **repetidos** cancelan algo **menos** que estos.

No hay gente con menos o iguales cancelaciones que no cancelaciones en su historial que haya cancelado ahora.

En **todos** estos grupos **dominan** las **no cancelaciones**.

Sector 1 cancela menos frente a los otros dos.

Proporción del estado actual de las cancelaciones



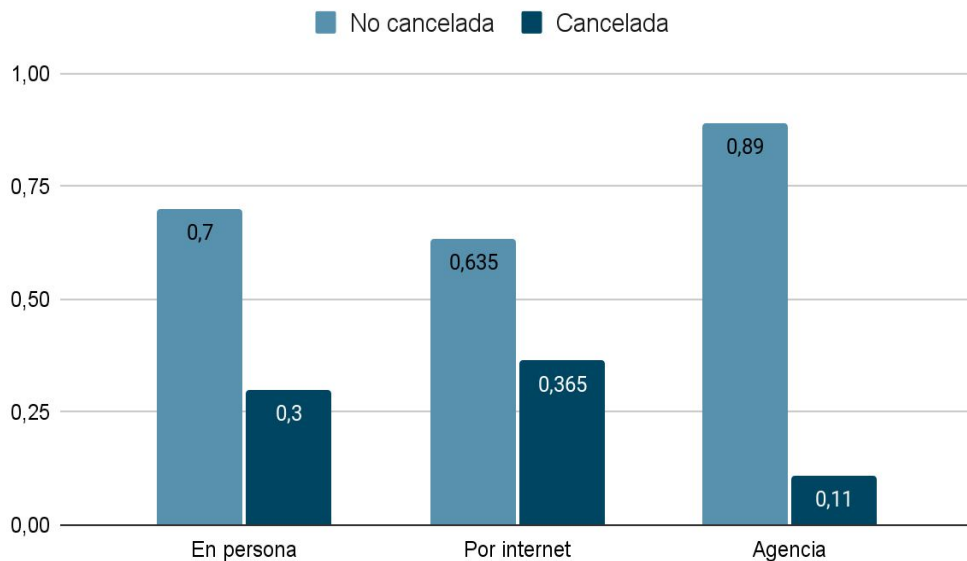
Aquí vemos la diferencia entre el **sector 1 y los otros dos.**

El **sector 2 y el sector 3** tienen las mismas **proporciones.**

Aún así, las **no cancelaciones** siguen **predominando.**

Las no cancelaciones también dominan según la forma de reservar.

Proporción del estado actual de las cancelaciones



En los **3 grupos** sigue siendo **mayoritario** el hecho de **no cancelar**.

Las reservas **en persona y por internet** tienen las **mismas proporciones**. En cambio las reservas por **agencia** tienen **menor** ratio de **cancelación**, lo cual tiene bastante sentido.

El resto de formas de reserva las omitimos por ser insuficiente cantidad.

Conclusiones:

- **Precios:**

- Los valores de los precios medios por habitación muy cercanos a 0 tienen relevancia y podrían ser errores en la toma de datos, se eliminan.
- Tiene cierto sentido juntar los tipos de habitación por pares.
- Separar por fechas no parece interesante.
- Relacionar con pensiones puede tener interés haciendo la separación en 2 grupos: los que no seleccionan y los que seleccionan alguno; e ignorar el tipo 3.
- Tiene interés separar con/sin niños.
- No es relevante la antelación de la reserva.

- **Familias:**

- La gran mayoría de clientes son adultos, así que será interesante centrarse en este público.
- Se puede buscar incentivar a las familias y mejorar esas ventas.
- No merece la pena separar entre ir con o sin niños para ver días de semana o parking, pero sí para requisitos especiales.

- **Cancelaciones:**

- En cualquier separación por grupos hecha aquí hay una gran mayoría de no cancelaciones, habrá que ver más con detalle en cuáles hay más tendencia.

Next steps:

- **ANOVA/MANOVA** para verificar si hay diferencias estadísticamente significativas entre diferentes grupos de interés y poder tomar decisiones fundamentadas estadísticamente.
- **Predecir** según nuestras variables qué clientes van a **cancelar** su reserva: de esta forma podremos clasificar clientes según sus datos y aplicar las medidas convenientes para disminuir pérdidas (fianzas, ofertas, etc.)

Backup

Summary:

<u>Booking_ID</u>	<u>no of adults</u>	<u>no of children</u>	<u>no of weekend nights</u>
Length:36275	Min. :0.000	Min. :0.0000	Min. :0.0000
Class :character	1st Qu.:2.000	1st Qu.:0.0000	1st Qu.:0.0000
Mode :character	Median :2.000	Median :0.0000	Median :1.0000
	Mean :1.845	Mean :0.1053	Mean :0.8107
	3rd Qu.:2.000	3rd Qu.:0.0000	3rd Qu.:2.0000
	Max. :4.000	Max. :10.0000	Max. :7.0000

<u>arrival date</u>	<u>market segment type</u>	<u>repeated quest</u>
Min. :1.0	Length:36275	Min. :0.00000
1st Qu.:8.0	Class :character	1st Qu.:0.00000
Median :16.0	Mode :character	Median :0.00000
Mean :15.6		Mean :0.02564
3rd Qu.:23.0		3rd Qu.:0.00000
Max. :31.0		Max. :1.00000

<u>no of week nights</u>	<u>type of meal plan</u>	<u>required car parking space</u>
Min. :0.000	Length:36275	Min. :0.00000
1st Qu.:1.000	Class :character	1st Qu.:0.00000
Median :2.000	Mode :character	Median :0.00000
Mean :2.204		Mean :0.03099
3rd Qu.:3.000		3rd Qu.:0.00000
Max. :17.000		Max. :1.00000

<u>no of previous cancellations</u>	<u>no of previous bookings not canceled</u>
Min. :0.00000	Min. :0.0000
1st Qu.:0.00000	1st Qu.:0.0000
Median :0.00000	Median :0.0000
Mean :0.02335	Mean :0.1534
3rd Qu.:0.00000	3rd Qu.:0.0000
Max. :13.00000	Max. :58.0000

<u>room type reserved</u>	<u>lead time</u>	<u>arrival year</u>	<u>arrival month</u>
Length:36275	Min. :0.00	Min. :2017	Min. :1.000
Class :character	1st Qu.:17.00	1st Qu.:2018	1st Qu.:5.000
Mode :character	Median :57.00	Median :2018	Median :8.000
	Mean :85.23	Mean :2018	Mean :7.424
	3rd Qu.:126.00	3rd Qu.:2018	3rd Qu.:10.000
	Max. :443.00	Max. :2018	Max. :12.000

<u>avg price per room</u>	<u>no of special requests</u>	<u>booking status</u>
Min. :0.00	Min. :0.0000	Length:36275
1st Qu.:80.30	1st Qu.:0.0000	Class :character
Median :99.45	Median :0.0000	Mode :character
Mean :103.42	Mean :0.6197	
3rd Qu.:120.00	3rd Qu.:1.0000	
Max. :540.00	Max. :5.0000	