TALLER 1- INFORME HOTELERO

Diego Felipe Carvajal Lombo - 201911910

Jesús Manuel Ospino Bernal - 201915195

En el siguiente informe, se presentan los hallazgos clave obtenidos del análisis de datos de reservas de hoteles. Esto con el objetivo de poder optimizar las estrategias de ocupación de las cadenas hoteleras, política de precios y gestión de cancelaciones. Para esto, se llevó a cabo un análisis de los datos para descubrir patrones y tendencias de las personas que realizan reservaciones en los hoteles.

1. Calidad, limpieza y exploración de datos

En primer lugar, se realizó una exploración de la calidad de los datos donde se corrigieron inconsistencias de los datos con la realidad del negocio. Por ejemplo, se eliminaron los registros los cuales contenían una cantidad negativa de personas.

Seguidamente, se utilizó la herramienta pandas profiling para obtener un reporte detallado de cada columna de los datos y poder identificar más inconsistencias. Con esto, encontramos que las variables company, agent y kids son aquellas con mayor cantidad de valores nulos. Es por tal razón que se iniciará la revisión y limpieza pertinente del dataset por estas variables.

En tercer lugar, se seleccionaron 5 variables que consideramos más importantes para poder realizar el análisis solicitado. Las cuales son las siguientes:

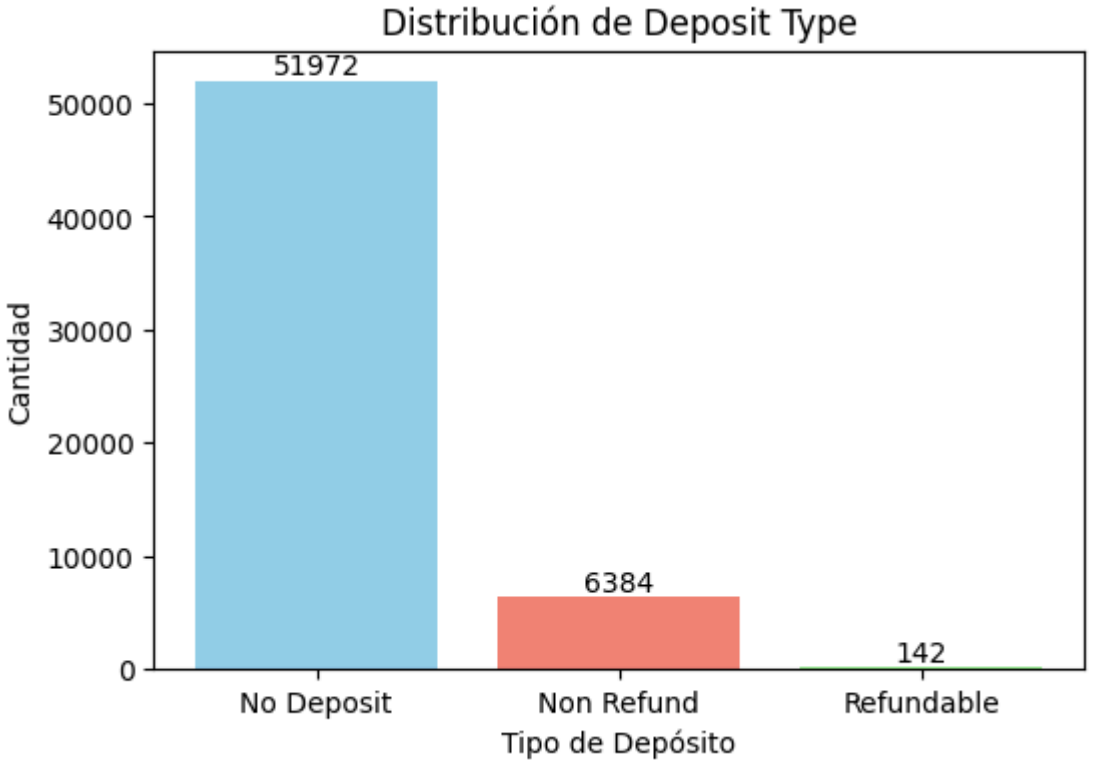
- Deposit Type
- ADR (Promedio gastado por noche)
- Is Canceled
- Lead Time
- Arrival Date Month\

Cabe resaltar que se calculó el número total de personas, sumando los valores de adultos, niños y bebés. Sin embargo, podemos ver que la gran mayoría (>90%) tiene únicamente 2 personas. Por lo que va a ser muy difícil

obtener insights de un valor tan constante. Por lo que decidimos no usar esta columna calculada.

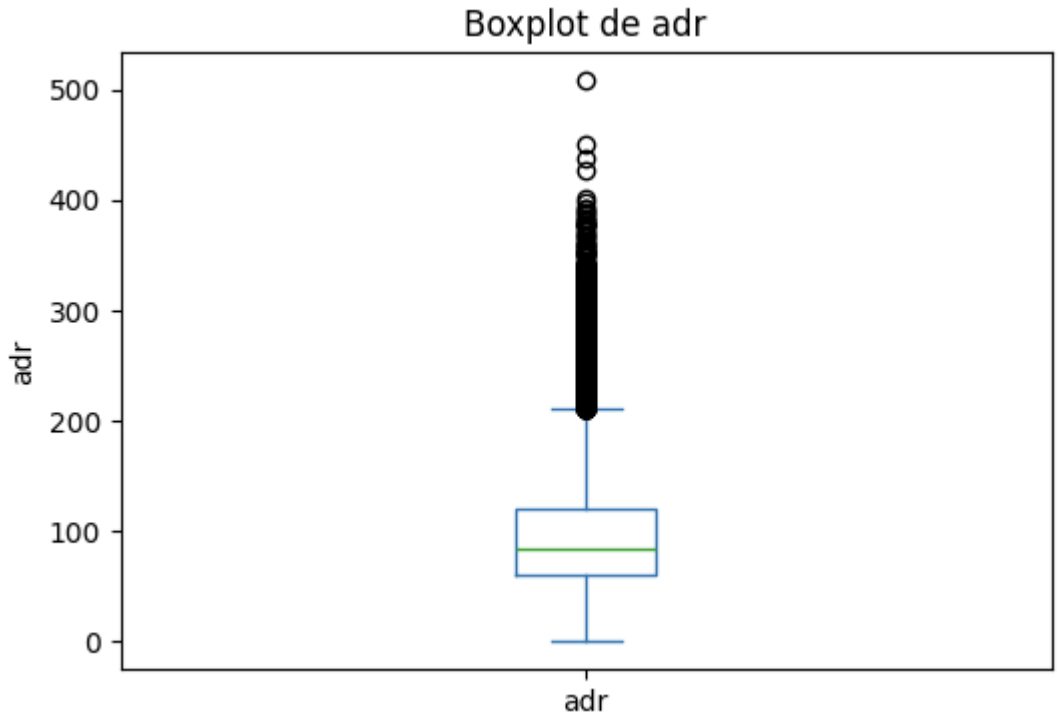
Deposit Type

Esta columna nos indica el tipo de depósito realizado por la persona. Esta tiene 3 tipos: Reembolsable, No reembolsable y Sin depósito. Al hacer la exploración de esta podemos ver que la mayor predominancia es que las personas no realizan un depósito al hacer la reserva. Sin embargo, puede ser interesante que aproximadamente el 20% de las reservas si tienen un tipo de depósito, y esto puede afectar si la persona realiza la cancelación del servicio.

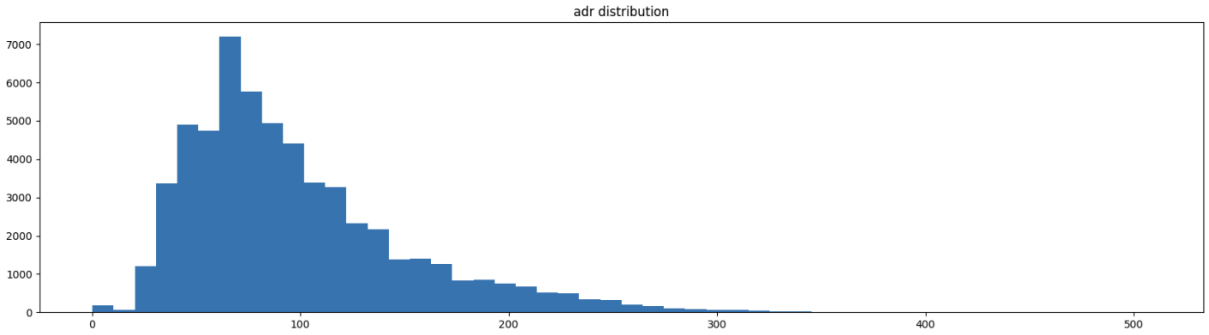


ADR

Esta columna nos da el valor promedio por noche que gastan las personas de una reserva. Es decir, toma el total del valor de las transacciones realizadas sobre el número de noches. Esta variable es clave en el análisis dado que es el único valor acerca del gasto de los residentes en los hoteles, al querer revisar la política de precios de estos, este valor nos va a dar la gran mayoría de insights para los hoteles.



En el anterior boxplot, ya podemos encontrar mayor detalle al eliminar el registro cuyo adr era de 5400. Con esto, podemos ver que los valores normalmente están entre 0 y 200 dólares por noche gastados. Además, los valores atípicos pueden llegar hasta 500 dólares.

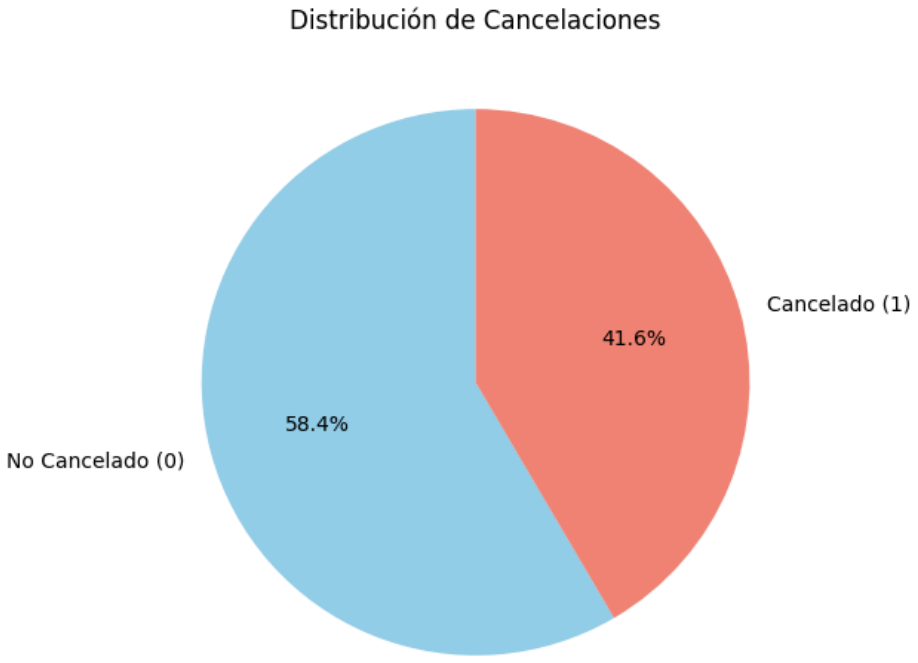


Podemos ver que los datos de adr no siguen una distribución de campana. Es decir, los datos están mayormente entre 20 y 70, a pesar de que su media es 97.

	adr
count	57575.000000
mean	97.775524
std	53.349827
min	0.260000
25%	61.000000
50%	85.000000
75%	121.000000
max	508.000000

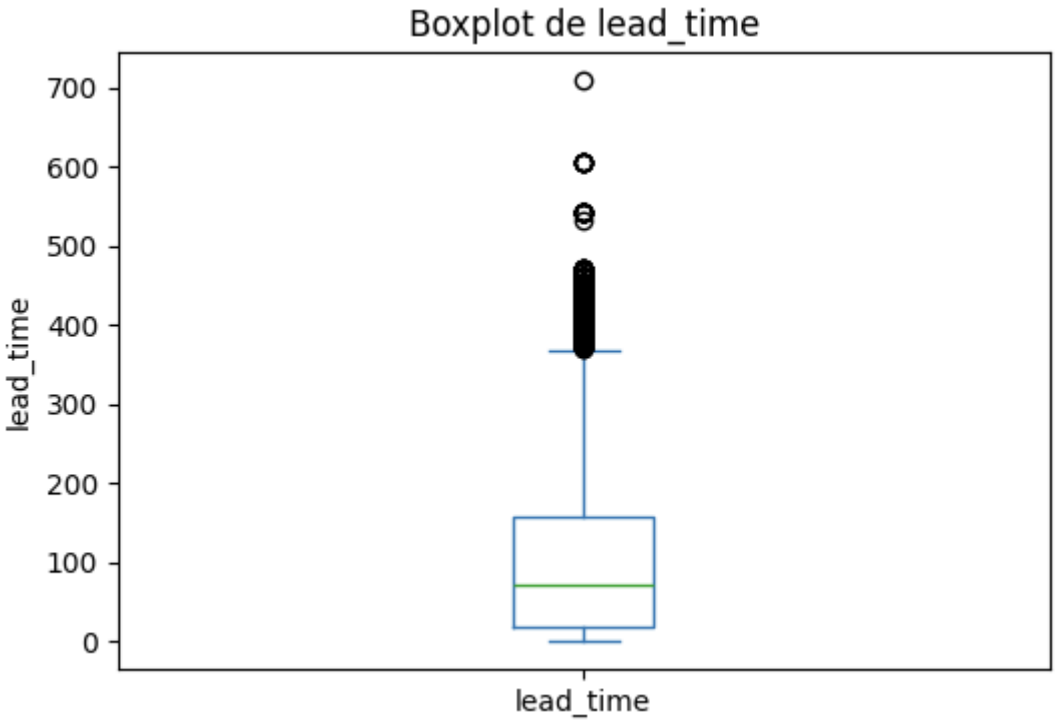
Is_Canceled

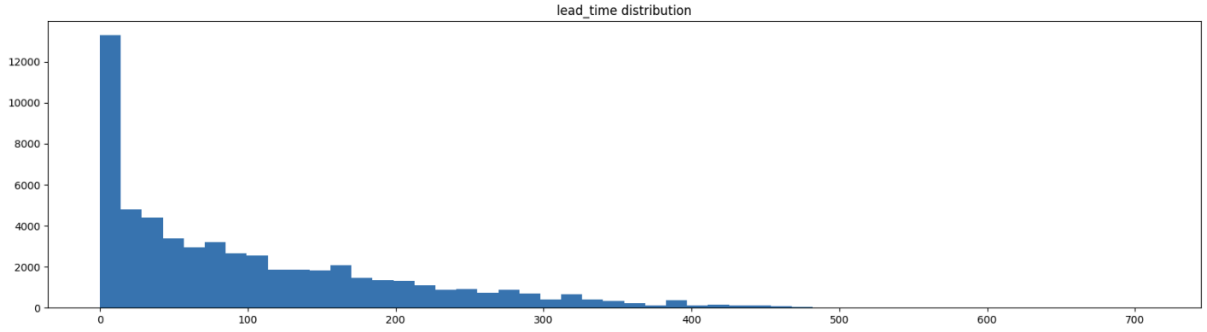
La variable is canceled nos indica si la reserva fue cancelada o no, esta es la más importante, dado que, es con la cual nos puede decir qué variables pueden hacer que se cancele o no una reserva. La variable es de tipo boolean. Se realizó un pie chart, que nos indica que el 58.4% de las reservas no fueron canceladas. Sin embargo, hay un gran porcentaje (41.6%) que si fueron cancelados, lo que perjudica a los hoteles para la planeación de los servicios prestados.



Lead_Time

Esta columna es el tiempo en días que pasa entre la reserva realizada y la llegada al hotel, en otras palabras, la anticipación con la que las personas realizan su reserva. Esta variable es importante para poder analizar el comportamiento de las personas, y si es posible que el tiempo afecte el dinero gastado o la cancelación de la misma. Del diagrama de caja, podemos visualizar que el lead time se encuentra normalmente distribuído entre 0 y 400, lo que es acorde al contexto del negocio, ya que consideramos que es normal que una persona realice una reserva con hasta un año de anticipación. También podemos encontrar que existen outliers con un valor mayor a 400. Estos se dividen en 2 grupos; entre 400 y 500 días y mayores a 500 días

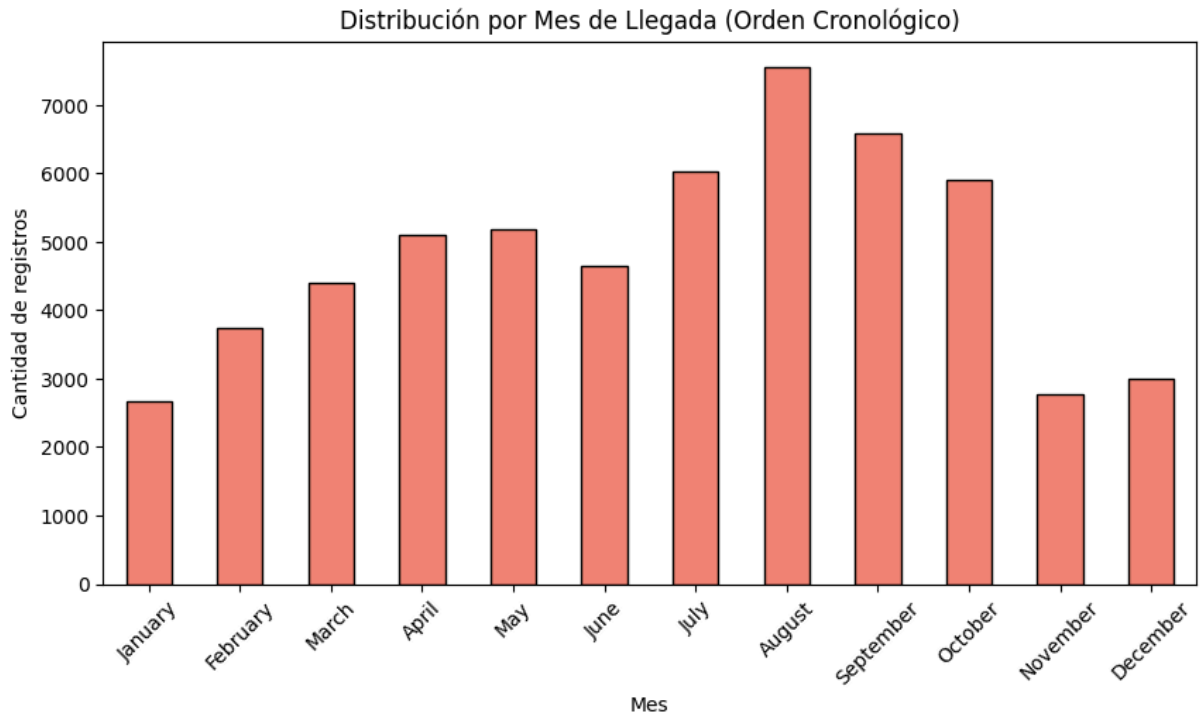




En la gráfica anterior, podemos ver que sigue el comportamiento de una distribución exponencial los valores de anticipación de la reserva, lo cual corresponde al negocio, dado que es más normal reservar con poco tiempo el hotel.

Arrival Date Month

Esta columna nos dice en qué mes está programada la reserva, es decir, el mes en el que llegan los turistas al hotel. Es de tipo categórica, y, naturalmente, cada categoría corresponde a un mes del año. Esta es importante dado que nos puede dar un indicio si la época del año o la fecha afecta las decisiones de las personas en cuanto al dinero a gastar y cancelar los servicios.



2. Estrategia de análisis

Para poder realizar el análisis correctamente, vamos a ver la relación y correlación que hay en las variables anteriormente mencionadas para ver si hay afectaciones o comportamientos específicos en las personas que más suelen gastar dinero y que suelen cancelar. Esto con el fin de que con estos

patrones los hoteles puedan reajustar sus políticas y aumentar las ganancias y gestiones.

Para esto, vamos a utilizar las técnicas como Pearson y Spearman para ver si existe una correlación entre las variables. También haremos cálculos sobre las columnas para generar nuevas columnas calculadas como tasa y porcentajes, esto con el fin de evitar sesgos de mayoría.

En general, las preguntas guías del desarrollo de este estudio serán las siguientes:

1. ¿Cuáles son los meses con mayor porcentaje de cancelación?
2. ¿En qué meses se gasta más dinero las personas en las reservas?
3. ¿El tipo de reserva afecta las cancelaciones?
4. ¿El tiempo de anticipación de la reserva puede ayudar a predecir si se va a cancelar o no?
5. ¿El tiempo de anticipación afecta la cantidad de dinero gastado por las personas?

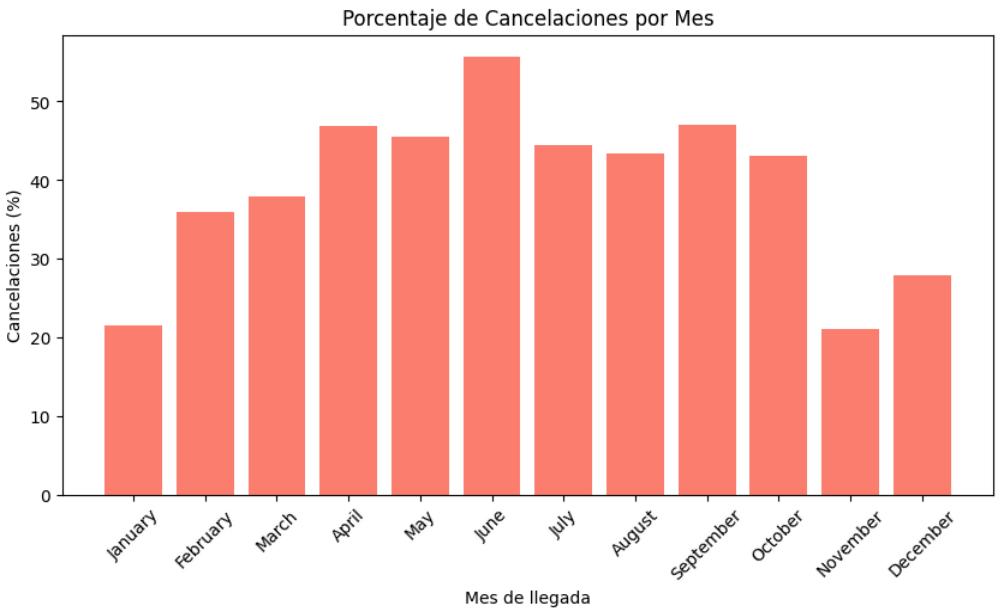
3. Desarrollo de la estrategia

Antes de iniciar el desarrollo de las preguntas, se hicieron pruebas de normalidad en las variables numéricas usando la prueba de sapphire.

El desarrollo se dio pregunta a pregunta:

3.1 ¿Cuáles son los meses con mayor porcentaje de cancelación?

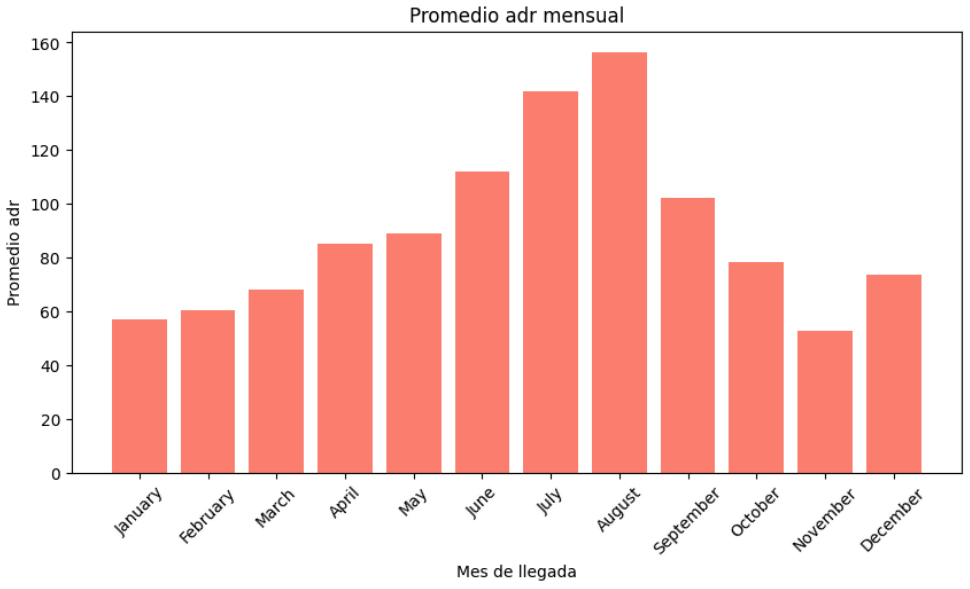
Con el fin de evitar sesgos relacionados con la frecuencia de los datos para cada de uno de los periodos, se tomó la decisión de realizar el cálculo del porcentaje de cancelaciones realizadas en cada uno de los meses, ya que existían meses como noviembre, diciembre y enero que tenían alrededor de 2500 reservaciones y otros meses como Julio y Agosto, en los cuales las reservaciones rondaban los 7000.



Una vez hecha la transformación, se evidencia que Junio es el mes con mayor tasa de cancelación, a pesar no ser uno de los meses con mayor cantidad de reservas, y que noviembre, diciembre y enero evidenciaban tasas de cancelación más bajas. Adicionalmente, se evidencia una tendencia a aumentar la tasa de cancelación conforme se acerca junio y a disminuir conforme se aleja.

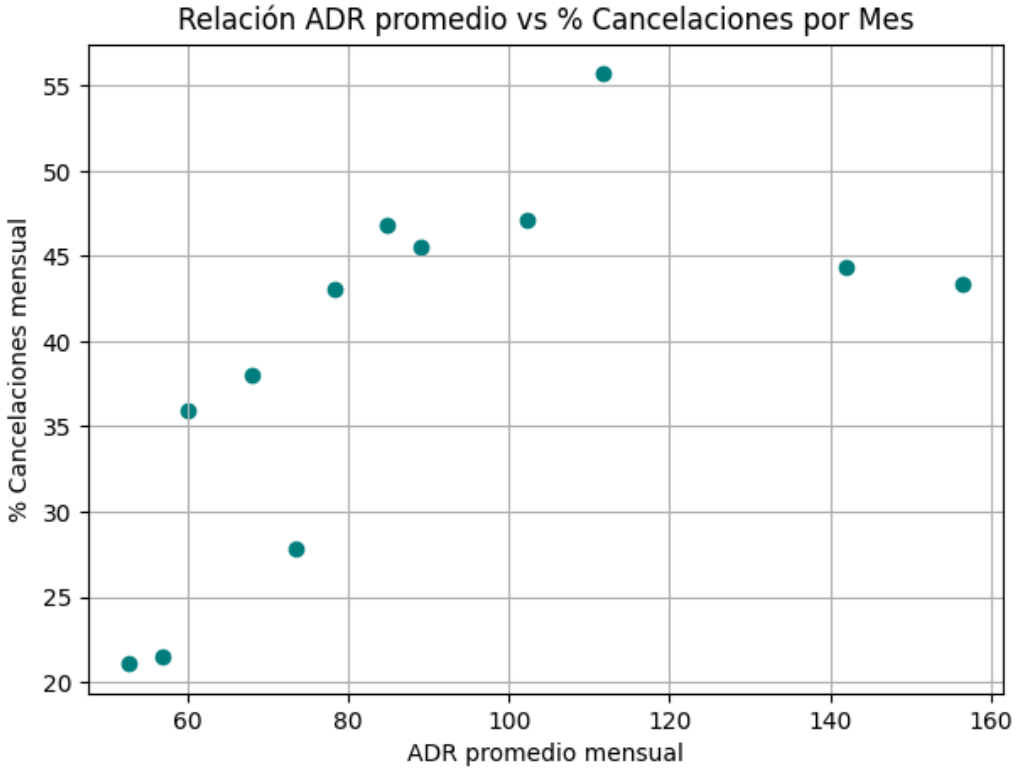
3.2 ¿En qué meses se gasta más dinero las personas en las reservas?

Para responder esta pregunta se realizó el cálculo de la media del adr agrupando por el mes de llegada de los usuarios al lugar de reservación.



De acuerdo al gráfico, obtuvimos que conforme el verano (junio-agosto) se acerca, existe un promedio de costo por noche más alto en la reservación, mientras que conforme se aleja, el costo medio por noche decrece. Lo cual tiene sentido con los periodos de temporada alta en la industria hotelera.

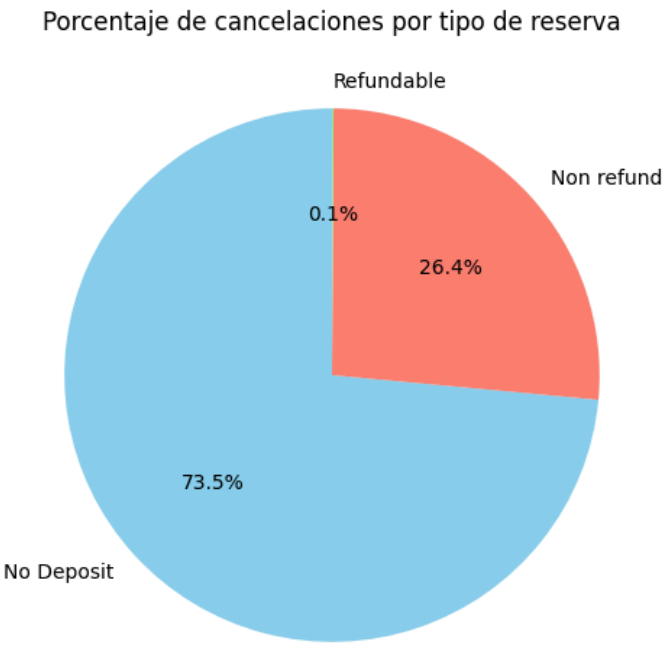
Finalmente, quisimos revisar si existía correlación entre la el la tasa de cancelaciones mensual y el promedio de adr mensual, que según la prueba de Spearman , si se encuentra significativamente correlación con un coeficiente de 0.7.



3.3 ¿El tipo de reserva afecta las cancelaciones?

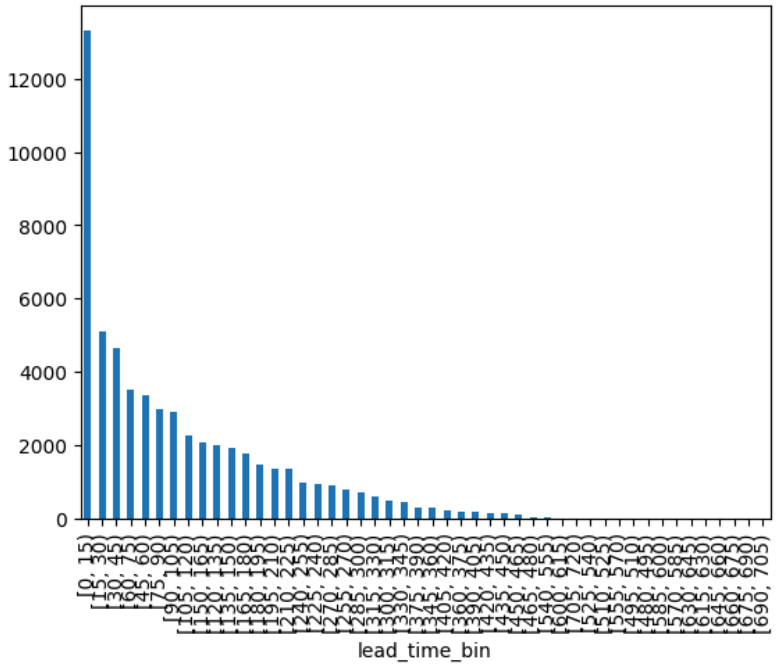
Para realizar este análisis, se debió calcular el porcentaje que representa cada tipo de reserva en las cancelaciones, para esto se sumó la cantidad de tipos y se dividió sobre el total de cancelaciones.

Los resultados mostraron que 3 de cada 4 cancelaciones son de tipo sin depósito, más exactamente el 73.5%. La otra mayoría se encuentra en las cancelaciones de tipo Sin devolución, con un 26.4% o una tasa de 1 de cada 4 aproximadamente. Por último, el menor porcentaje de cancelaciones en de tipo Con Devolución, representando únicamente el 0.1%.

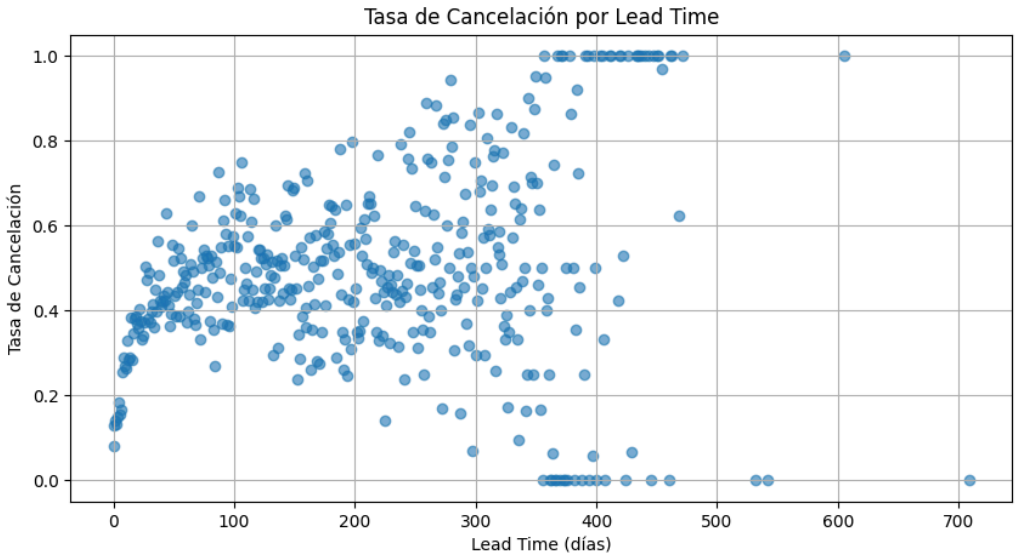


3.4 ¿El tiempo de anticipación de la reserva puede ayudar a predecir si se va a cancelar o no?

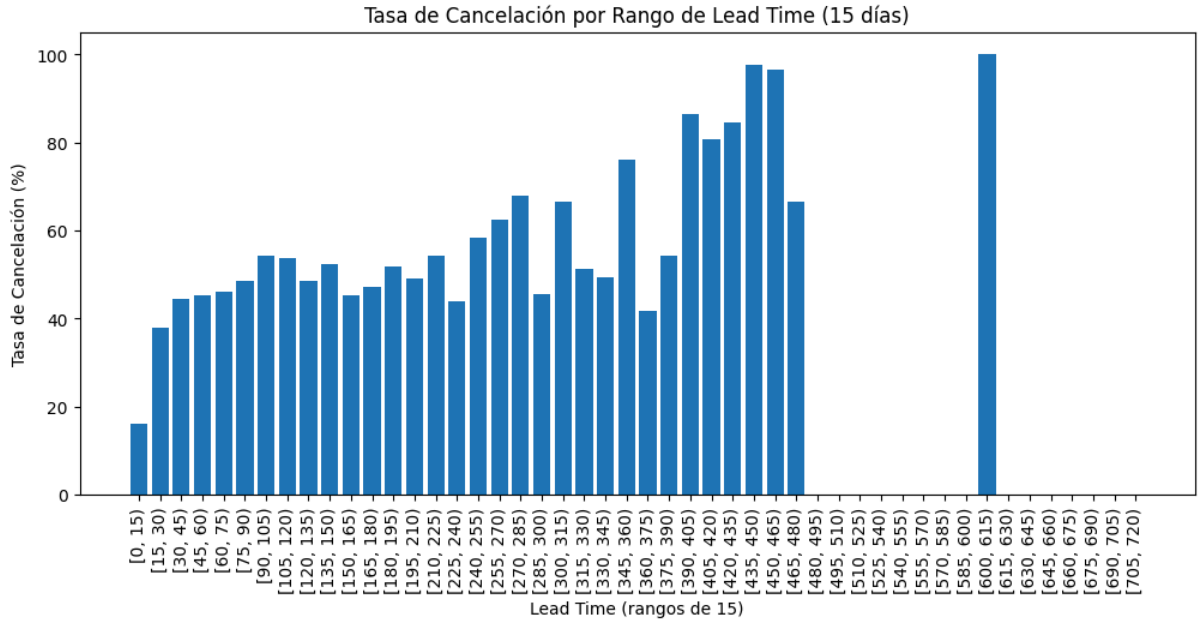
Ahora bien, para darle respuesta a esta pregunta primero realizamos el cálculo de la tasa de cancelaciones por la cantidad de días que habían transcurrido entre el día de agendamiento y el día de llegada al lugar. Esto, para evitar el sesgo de la frecuencia que se pueda dar en esta relación. Lo anterior, ya que como se evidencia en el siguiente gráfico:



Conforme el tiempo de entre reserva y día de reserva se hace más grande, la cantidad de reservas con ese comportamiento disminuye. Motivados, visualizamos la relación:



De la cual a simple vista no evidenciamos una relación clara en los datos. Por otro lado, dado que existe una relación monótona decreciente entre las dos variables, como se evidencia en el primer gráfico, calculamos el coeficiente de Spearman y efectivamente existe una fuerte correlación negativa entre el lead time y la cantidad de cancelaciones. Sin embargo, como este comportamiento no parecía estar representado en la segunda gráfica, decidimos tomar rangos de tiempo de 15 días y calcular entre estos rangos de tiempo cuántas eran las cancelaciones. Al hacerlo, obtuvimos el siguiente gráfico:

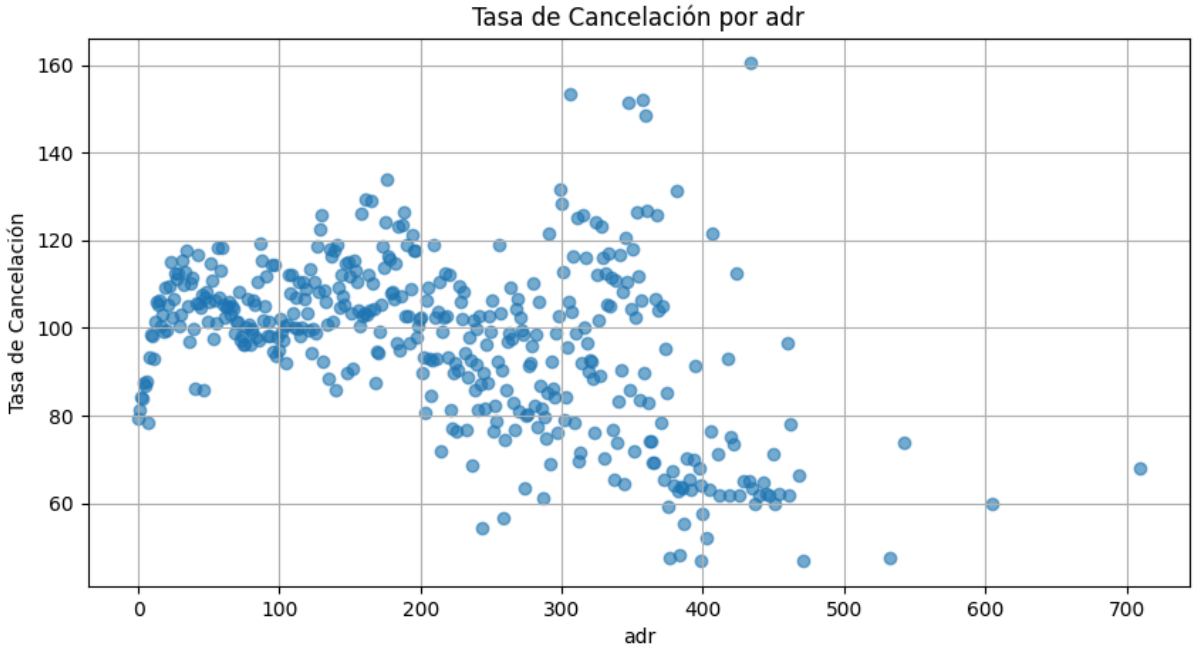


En cual se evidencia una correlación monótona positiva entre el rango de tiempo y la tasa de cancelación. Como evidenciamos que la data se agrupa entre 0 y 500, el dato del rango de 600 lo tomaremos como un dato atípico, y por lo tanto, para realizar una prueba de correlación no lo tendremos en cuenta. Una vez realizada la prueba, se encontró que como el coeficiente de Spearman es positivo y de 0.7, se confirma una correlación monota positiva y fuerte entre estas dos variables.

3.5 ¿El tiempo de anticipación afecta la cantidad de dinero gastado por las personas?

Al igual que se realizó previamente, para poder responder esta pregunta, se hizo una conversión de los rangos (categorías) en códigos ordinales. Con esto calculado, se procedió a hacer el cálculo de spearman para saber si es posible encontrar una correlación entre las variables de ADR y lead_time.

Luego de realizar esto, se observó que se obtuvo un coeficiente de -0.392. Lo que nos indica que estas dos variables no están fuertemente correlacionadas. Esto nos indicó que no hay ninguna afectación entre el tiempo de anticipación de la reserva con la cantidad de dinero gastado.



4. Resultados

Según el análisis realizado previamente, se pueden obtener las siguientes conclusiones, y con ellos las siguientes estrategias a la cadena hotelera para cumplir con sus objetivos:

1. La mayoría de las cancelaciones se realizan en la mitad del año natural. Es decir, cerca al mes de Junio, aumentan cerca de este mes y disminuyen conforme se alejan.
   Con base a esto, se recomienda a la cadena hotelera hacer un presupuesto de recursos contando con las reservas sobre los meses de mayo y junio, ya que tienen una alta probabilidad de ser cancelados.

2. Se pudo evidenciar que los meses donde más dinero se gastan las personas es en el verano, es decir, los meses de Julio y Agosto. Lo que implica que hay una alta disposición a pagar en esos meses.

   Con respecto a este insight, se le recomienda a la cadena hotelera hacer un sistema de tarifa dinámica, que incremente el precio de sus servicios a medida que llegan estos meses, y disminuya cuando ya han pasado.

   También es posible ofrecer tarifas premium o paquetes exclusivos (ej. habitaciones con vista, acceso a spa, cenas gourmet) en estos meses, ya que los clientes están dispuestos a gastar más.

   Por último, se pueden crear paquetes familiares y de ocio (ej. “summer experience”) que incluyan más de un servicio y aumenten el gasto promedio por cliente.

3. 3 de cada 4 reservas son de tipo Sin depósito.

   Dado este alto porcentaje de cancelaciones que corresponden al tipo sin depósito, se recomienda a la cadena hotelera disminuir la cantidad de reservas que se pueden hacer bajo esta modalidad, puede ser, disminuyendo la cantidad de habitaciones que la tienen.

   Otra forma puede ser implementar un plan de cancelación gratuito únicamente durante un tiempo definido, por ejemplo, 7 días, luego de

esto puede haber un esquema de cancelación dependiendo de la cantidad de tiempo faltante para la reserva.

4. La cantidad de días de anticipación está correlacionado con la tasa de cancelación de las reservas, es decir, conforme es más amplio el tiempo de anticipación, mayor es la tasa de cancelación.

   Por tal razón, se le aconseja a las cadenas hoteleras que tengan un parámetro de tiempo no mayor a un año para hacer las reservas, y con esto disminuir la probabilidad de cancelación.

5. La cantidad de tiempo de anticipación de la reserva no afecta el dinero gastado por las personas.

   Debido a esto, no se le recomienda a los hoteles hacer ningún tipo de estrategia o marketing basado en aumentar o disminuir la cantidad de tiempo de anticipación de la reserva con el fin de obtener más ganancias directas, ya que no se observa ninguna correlación.