Daniela Camacho Molano - 202110974
Laura Ruales - 201617340

Taller 1 Demanda de ocupación hotelera

Objetivo: Realizar un análisis de datos utilizando técnics estadísticas y de visualización que permitan entender el conjunto de datos para descubrir insights

Entendimiento inicial de datos

El dataset *Hotel Bookings* contiene 58895 registros y 33 atributos. Los cuales se dividen entre 17 variables categóricas (Tipo de hotel, cancelación, tipo de comida, país de origen, segmento de mercado, canal de distribución, huésped anteriormente, cancelaciones previas, reservas previas no canceladas, tipo de habitación reservada, tipo de habitación asignada, tipo de depósito, agente, compañía, tipo de cliente, estado de la reserva) y 16 variables cualitativas (días entre la fecha de reserva y de llegada, fecha de llegada dividida entre semana, día, mes y año, noches de reserva entre semana y fines de semana, cantidad de niños, adultos y bebes, cantidad de cambios a la reserva, días en lista de espera, tarifa diaria promedio, espacios de parqueadero solicitados, total de solicitudes especiales, fecha de actualización de estado de la reserva).

Después de realizar la identificación y exploración de las variables, tanto categóricas como numéricas. Se realizo la identificación de la variable que representa la cancelación de una reserva, esta se denomina "is_canceled", la cual es boolena. Dado el contexto del negocio, esta variable se define como la variable objetivo para describir y analizar su comportamiento en función de otras variables relevantes.

Dado esto, se realizaron graficas con ciertas variables que dado el contexto del negocio se tiene como hipótesis inicial que pueden tener una relación con la variable objetivo, estas graficas tienen como fin validar de manera inicial la hipótesis realizada. De esta forma, se escogen 5 variables que son considerados importantes para el análisis a realizar.

1. Variable Fecha:

Es una manipulación de las variables arrival_date _year y arrival_date_month que son variables numéricas y representan cada una el año y el mes de la reserva. La variable arrival_date _year tenía valores como "20016" que por el contexto se remplazaron con 2016. De este modo se construyó la variable de Fecha de reserva que comprende un periodo de 2 años entre Julio-2015 y Julio 2017.

Inicialmente, se hizo un análisis gráfico de la cantidad de reservas por mes, así como la correspondiente tasa de cancelación. En donde se encontró que el año 2015 tiene un pico en cantidad de reservas en el mes de septiembre. Mientras que el año 2016 se mantiene una cantidad de reservas más o menos constante entre marzo y octubre. Sin embargo, el crecimiento de la cantidad de reservas para este año se ve acompañado del aumento de la tasa de cancelación. Y finalmente en el año 2017 tanto la cantidad de reservas como la tasa de cancelación es menor. Tal como se puede observar en la gráfica 1.a

Por otro lado, en la gráfica 1.b se puede observar que en el año 2016 la cantidad de reservas canceladas superaron la cantidad de reservas sin cancelar. Además, que el año 2015 tiene un pico de reservas efectivas (sin cancelar) en octubre. Pero en el resto del periodo considerado la cantidad de reservas efectivas es aproximadamente constante.

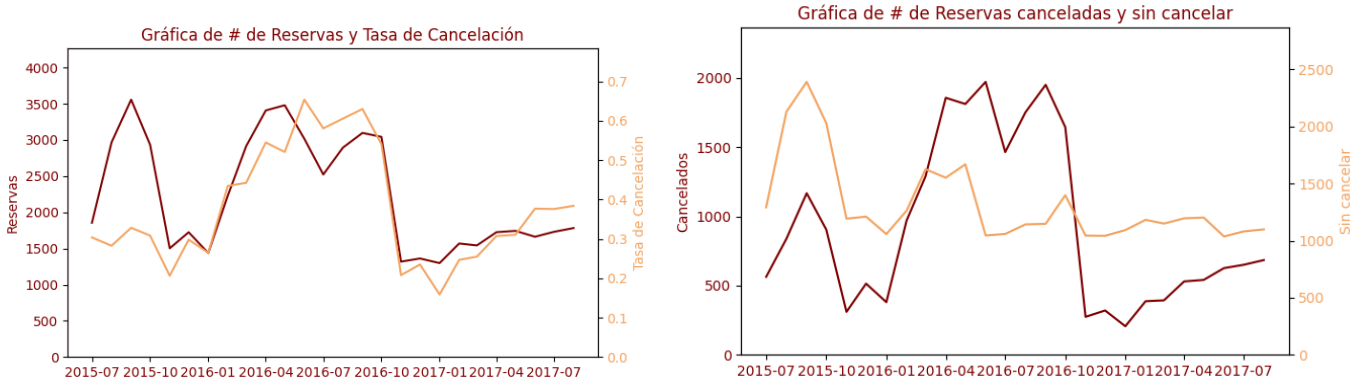


Figura 1. a. Número total de reservas y Tasa de Cancelación por mes.

b. Número de reservas No Canceladas y Canceladas por mes

Por otro lado, también modificamos la presentación de los registros para analizar estacionalidad según la época del año. Comparando la cantidad total de Reservas y la tasa de cancelación en el mismo periodo de cada año. En los años 2015 y 2016 se observa una reducción del número de reservas en el mes de noviembre que se sostiene hasta enero del año siguiente. Así mismo, hay un pico de reservas al rededor del mes de septiembre. Por otra parte, en cuanto al a tasa de cancelación tanto en el año 205 como 2017 les aproximadamente constante mientras que en el año 2016 se presentan aumentos y reducciones dramáticas. Sin embargo, no se cuenta con suficientes datos para notar un patrón.

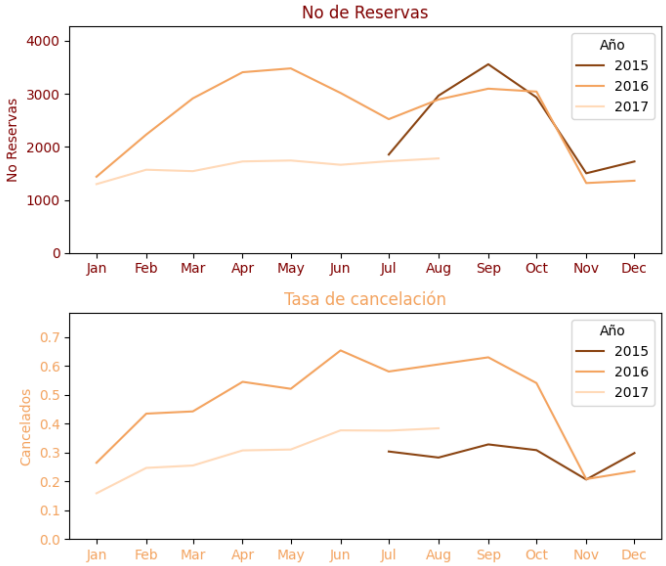


Figura 2. Número total de reservas y Tasa de Cancelación por mes.

2. Variable Hotel:

La variable *hotel* es de tipo categórica y representa el tipo de hotel asociado a cada reserva. Según el diccionario de datos, puede tomar 2 valores: *City Hotel* y *Resort Hotel.*

A partir de estas graficas, se identificó que la mayoría de las reservas se realizaron en un hotel de tipo Resort (73%). Mientras que sólo el 26% de las reservas se realizaron en un hotel de Ciudad. Por otro lado, la tasa de cancelación para los hoteles de Ciudad es 70% mas del doble de la tasa de cancelación en los hoteles tipo Resort (27%).

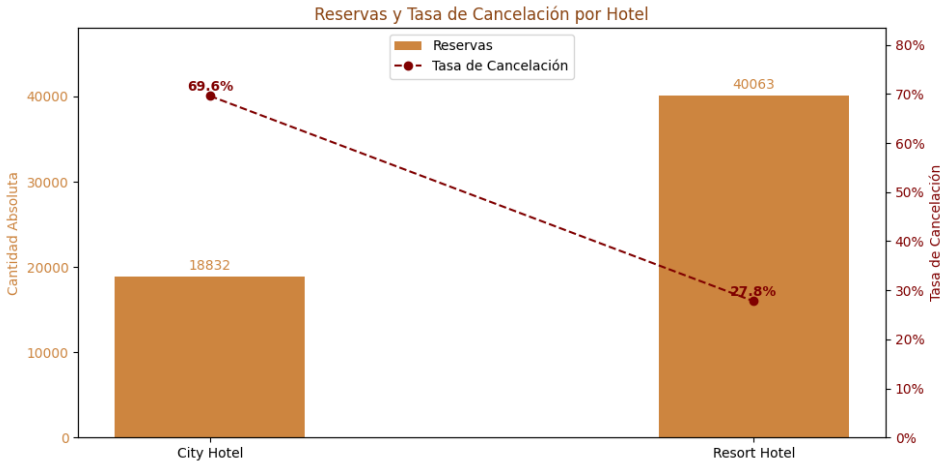


Figura 3. Número total de reservas y Tasa de Cancelación por Tipo de Hotel.

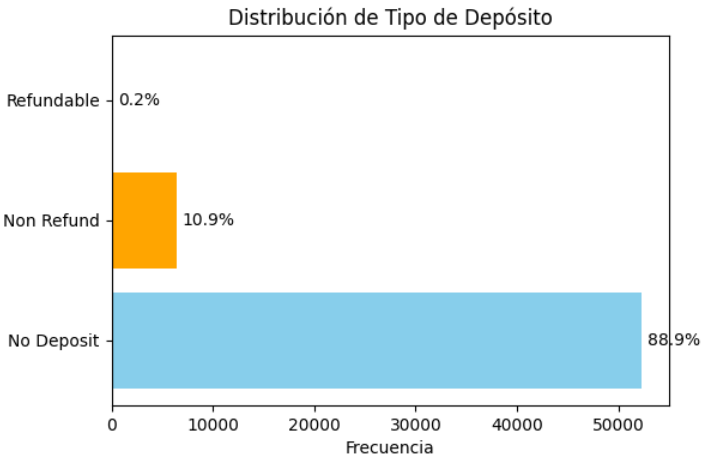
3. Variable Deposit Type:

La variable Deposit Type es de tipo categórica y representa el tipo de depósito asociado a cada reserva. Según el diccionario de datos, puede tomar tres valores: Refundable, Non Refund y No Deposit.

Durante la preparación de datos fue necesario unificar categorías, ya que se identificó duplicidad en la codificación: existían los valores "Non Refund" y "No Refund", que correspondían al mismo concepto. Estos se integraron bajo una única categoría ("Non Refund"), obteniendo así las tres categorías correctas definidas por el diccionario.

Posteriormente, se planteó la hipótesis de que las reservas clasificadas como No Deposit y Refundable presentan una mayor probabilidad de cancelación, dado que los huéspedes no asumen una pérdida económica significativa al cancelar.

Finalmente, se llevó a cabo un análisis univariado con el fin de comprender la distribución de la variable y el comportamiento de cada una de sus categorías. A partir de este análisis se observó lo siguiente:



A partir de estas graficas, se identificó los siguientes puntos:

- La mayoría de las reservas se realizaron bajo la modalidad No Deposit (88.9%), lo que evidencia una fuerte preferencia por opciones sin pago anticipado.

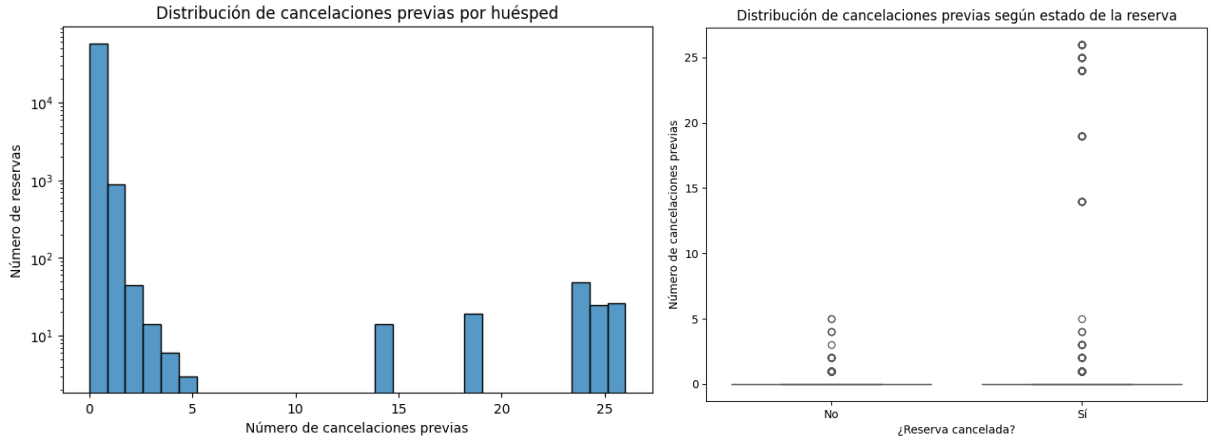
- Un 10.9% de las reservas son Non Refund, lo que indica que existe un grupo de clientes que acepta pagar por adelantado sin posibilidad de devolución, posiblemente a cambio de tarifas más bajas.
- La categoría Refundable apenas alcanza el 0.2%, mostrando que esta modalidad casi no es utilizada, ya sea porque no se ofrece con frecuencia o porque no resulta atractiva para los clientes.
- La distribución refleja que el negocio depende en gran medida de las reservas No Deposit, lo cual puede estar asociado a mayores tasas de cancelación, dado que los clientes no enfrentan pérdidas económicas al anular la reserva.

4. Variable Previous Cancellations:

La variable número de cancelaciones previas es de tipo numérica y refleja cuántas veces un huésped había cancelado antes de la reserva actual. Su rango de valores va de 0 a 26. En este caso no fue necesario realizar ningún tipo de preparación de datos.

A partir de esta variable se planteó la hipótesis de que los huéspedes con un mayor número de cancelaciones previas tienen una mayor probabilidad de cancelar nuevamente en comparación con aquellos que registran pocas o ninguna cancelación previa.

Para explorar esta idea, se llevó a cabo un análisis univariado acompañado de distintas gráficas que permiten observar mejor la distribución y el comportamiento de la variable.



A partir de estas gráficas se identificó los siguientes puntos:

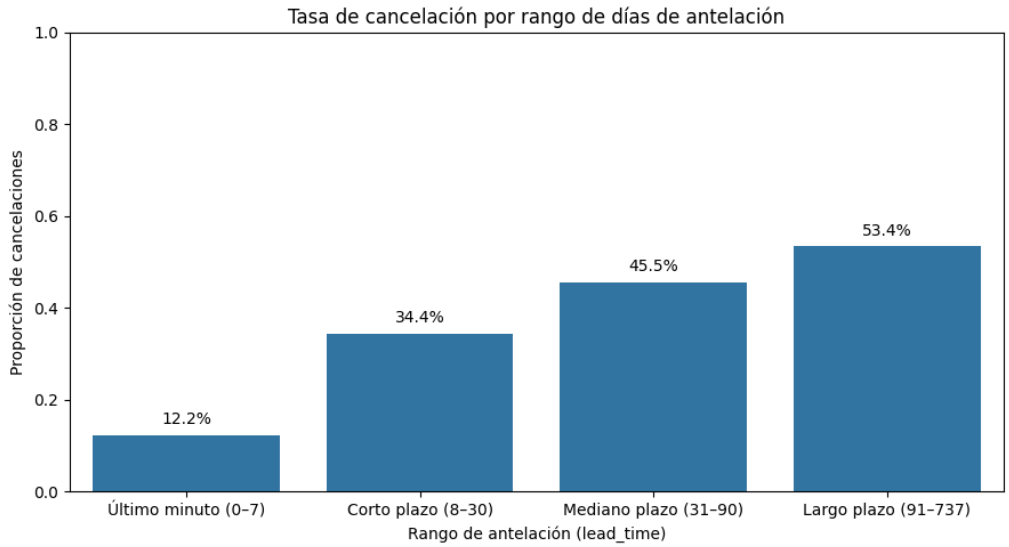
- La mayoría de los huéspedes registra cero cancelaciones previas, lo que indica que, en general, los clientes no tienen historial de cancelación.
- A medida que aumenta el número de cancelaciones previas, la frecuencia de huéspedes disminuye de manera considerable, mostrando que solo unos pocos acumulan varias cancelaciones.
- Se identifican valores atípicos con hasta 26 cancelaciones previas, lo que refleja la existencia de un pequeño grupo de clientes con un comportamiento muy distinto al promedio.
- Al comparar el estado de la reserva actual, se observa que los huéspedes con mayor número de cancelaciones previas aparecen con más frecuencia en el grupo de reservas canceladas, lo que respalda la hipótesis de que existe una relación entre el historial de cancelaciones y la probabilidad de cancelar nuevamente.
- La distribución de la variable es altamente sesgada a la derecha, concentrándose en valores bajos y con pocos casos extremos en los valores altos.

5. Variable Lead Time:

La variable Lead Time es de tipo numérica y representa el número de días entre la fecha en que se realizó la reserva y la fecha de llegada al hotel. Sus valores se encuentran en el rango de 0 a 737 días.

Se plantea la hipótesis de que, a mayor número de días de anticipación, existe una mayor probabilidad de que la reserva sea cancelada, ya que un periodo de tiempo más amplio incrementa la posibilidad de que surjan imprevistos o cambios en los planes del huésped.

Con el fin de comprender mejor el comportamiento de esta variable, se llevó a cabo un análisis exploratorio que permitió observar su distribución y posibles patrones relacionados con la cancelación de reservas.



- Se observa una clara relación entre el número de días de anticipación y la probabilidad de cancelación: a mayor lead time, mayor tasa de cancelación.
- Las reservas hechas a último minuto (0–7 días) tienen la menor tasa de cancelación (12.2%), lo que indica que, cuando la llegada es inminente, los huéspedes casi siempre cumplen con su reserva.
- En el rango de corto plazo (8–30 días) la tasa de cancelación aumenta significativamente a 34.4%, mostrando que con más días de margen aparecen más cancelaciones.
- En el mediano plazo (31–90 días) la tasa sigue creciendo hasta 45.5%, lo que refuerza la hipótesis de que el riesgo de cancelación aumenta con el tiempo de anticipación.
- Finalmente, en el rango de largo plazo (91–737 días) se alcanza la tasa más alta, con 53.4% de cancelaciones, es decir, más de la mitad de estas reservas no se concretan.
- El análisis confirma la hipótesis inicial: reservas realizadas con mayor anticipación están más expuestas a cancelarse, probablemente porque los planes de los clientes cambian más fácilmente cuando se hacen con mucho tiempo de antelación.

Estrategia de análisis

Como estrategia de análisis, se definió que para cada variable seleccionada se planteara una pregunta de negocio. Posteriormente, mediante el uso de pruebas estadísticas y técnicas de visualización, se buscará obtener insights relevantes que permitan comprender mejor la cancelación de reservas.

1. Fecha

Para esta variable se hizo una nueva modificación, se tomó solo el trimestre del año por lo que un mismo mes para diferentes años se considera dentro del mismo grupo. Además, no se consideró un trimestre fiscal tradicional sino un grupo de trimestres que inician de noviembre a enero puesto que coincide con un pico de demanda relacionado con las vacaciones de final de año. Finalmente, se planteó la siguiente pregunta de negocio:

¿El trimestre del año influye significativamente en la probabilidad de que una reserva sea cancelada?

A partir de la pregunta, se plantean las siguientes hipótesis:

- H0 (nula): La tasa de cancelación es igual independientemente del trimestre del año.
- H1 (alternativa): La tasa de cancelación depende del trimestre del año

Con un nivel de significancia: $\alpha = 0.05$

Dado que se busca comparar las proporciones de cancelación entre más de dos grupos, se utilizará una prueba chi cuadrado de independencia. La decisión se toma bajo el siguiente criterio:

- Si $p \leq \alpha$, se rechaza H0 - La tasa de cancelación depende del tipo de depósito.

2. Hotel

Para esa variable se planteó la siguiente pregunta de negocio:

¿El tipo de Hotel influye significativamente en la probabilidad de que una reserva sea cancelada?

A partir de la pregunta, se plantean las siguientes hipótesis:

- H0 (nula): La tasa de cancelación es igual independientemente del tipo de Hotel.
- H1 (alternativa): La tasa de cancelación depende del tipo de hotel.

Con un nivel de significancia: $\alpha = 0.05$

Dado que se busca comparar las proporciones de cancelación entre más de dos grupos, se utilizará una prueba chi cuadrado de independencia. La decisión se toma bajo el siguiente criterio:

- Si $p \leq \alpha$, se rechaza H0 - La tasa de cancelación depende del tipo de depósito.

3. *Deposit_type*

Para esa variable se planteó la siguiente pregunta de negocio:

¿El tipo de depósito influye significativamente en la probabilidad de que una reserva sea cancelada?

A partir de la pregunta, se plantean las siguientes hipótesis:

- H0 (nula): La tasa de cancelación es igual independientemente del depósito.
- H1 (alternativa): La tasa de cancelación depende del tipo de depósito.

Con un nivel de significancia: $\alpha = 0.05$

Dado que se busca comparar las proporciones de cancelación entre más de dos grupos, se utilizará una prueba chi cuadrado de independencia. La decisión se toma bajo el siguiente criterio:

- Si $p \leq \alpha$, se rechaza H0 - La tasa de cancelación depende del tipo de depósito.

4. Previous_cancellations

Esta es una variable numérica que representa el número de canacleciones hechas por el huespued previo a la reserva actual,en el data set actual el rango de valores es de [0,26]. Para esta variable se planteo la siguiente pregunta de negocio:

¿El número de cancelaciones previas de un cliente influye en la probabilidad de cancelar la reserva actual?

A partir de esto, se plantearon las siguientes hipótesis:

- H0 (nula): El promedio de cancelaciones previas es igual entre clientes que cancelan su reserva actual y los que no cancelan.

- H1 (alternativa): El promedio de cancelaciones previas es diferente entre clientes que cancelan su reserva actual y los que no la cancelan.

  Con un nivel de significancia: $\alpha = 0.05$

  Para este análisis se aplicó una prueba t de Welch, dado que permite comparar la media de cancelaciones previas entre dos grupos independientes: reservas canceladas y no canceladas. Esta prueba resulta adecuada porque la variable en estudio es numérica y, a diferencia del t-test clásico, no requiere asumir igualdad de varianzas, lo que lo hace más robusto ante posibles diferencias en la dispersión y tamaño de los grupos.

- Si $p \leq \alpha$, se rechaza H0 - el número de cancelaciones previas influye significativamente en la probabilidad de cancelar la reserva actual.

5. Lead_time

Esta es una variable numérica que indica el número de dias que hay entre el día que se realiza la reserva y el día de llegada al hotel; el rango de valores es de [0,737]. Para esta variable se planteó la siguiente pregunta de negocio:

¿Reservar con más días de antelación aumenta la probabilidad de que la reserva sea cancelada?

- H0: El tiempo de antelación de la reserva (lead_time) no está relacionado con la probabilidad de cancelación.
- H1: A mayor tiempo de antelación, mayor es la probabilidad de que la reserva sea cancelada.

Para este análisis se emplea una prueba t de Welch, ya que permite comparar el promedio de días de antelación (lead_time) entre reservas canceladas y no canceladas. Esta prueba es apropiada porque la variable es numérica, los grupos son independientes y, al no asumir varianzas iguales, resulta más robusto frente a diferencias en la dispersión y tamaño de las muestras, garantizando conclusiones más fiables.

Desarrollo de la estrategia

Dado las preguntas de negocio generadas en el punto anterior, se realizó una implementación en *jupyter notebook,* el cual se encuentra dentro del repositorio. En este se aplicaron las técnicas estadísticas y de visualización definidas para las 5 variables.

Generación de resultados

Dado el desarrollo de la estrategia mencionada en puntos anteriores, se va a realizar un análisis de los resultados obtenidos.

En primer lugar, se va a analizar los resultados de la variable

Fecha Se basó en los trimestres generados anteriormente y se utilizó una prueba de independencia de chi-cuadrado, se obtuvo un p-value = 0, valor inferior al nivel de significancia establecido ($\alpha = 0.05$). Esto permite rechazar la hipótesis nula (H0) y concluir que fecha tiene una influencia significativa en la probabilidad de cancelación de las reservas.

El tamaño del efecto calculado mediante Cramér's V = 0.161 se clasifica como pequeño, lo que indica que la relación entre ambas variables es estadísticamente significativa, pero desde una perspectiva práctica esto es poco relevante. Esto puede deberse a que la tasa de cancelación tiene una menor varianza intergrupal entre los trimestres 2, 3 y 4.

Por lo que se recomienda:

- Incorporar esta la temporada del año en los modelos de predicción de ocupación, de manera que se puedan anticipar cancelaciones y ajustar estrategias de overbooking para optimizar la gestión de ingresos durante los meses de mayor cancelación.
- Realizar campañas de incentivación para generar más reservas durante los meses noviembre, diciembre y enero

Por otro lado, para ***Hotel*** se utilizó una prueba de independencia de chi-cuadrado, se obtuvo un p-value = 0, valor inferior al nivel de significancia establecido ($\alpha = 0.05$). Esto permite rechazar la hipótesis nula (H0) y concluir que fecha tiene una influencia significativa en la probabilidad de cancelación de las reservas.

El tamaño del efecto calculado mediante Cramér's V = 0.396 se clasifica como pequeño, lo que indica que la relación entre ambas variables es estadísticamente significativa, y además desde una perspectiva práctica esto es relevante.

En este caso las recomendaciones para el negocio son

- Realizar campañas de incentivación para disminuir las cancelaciones en hoteles de tipo ciudad
- Incorporar el tipo de hotel en los modelos de predicción de ocupación, de manera que se puedan anticipar cancelaciones y ajustar estrategias de overbooking para optimizar la gestión de ingresos durante los meses de mayor cancelación.
- Priorizar la inversión en hoteles de tipo Resort sobre los de tipo Ciudad

Asi mismo, el análisis de la variable ***deposit_type*** se basó en el test de independencia chi-cuadrado, se obtuvo un p-value = 0, valor inferior al nivel de significancia establecido ($\alpha = 0.05$). Esto permite rechazar la hipótesis nula (H0) y concluir que el tipo de depósito tiene una influencia significativa en la probabilidad de cancelación de las reservas.

El tamaño del efecto calculado mediante Cramér's V = 0.411 se clasifica como grande, lo que indica que la relación entre ambas variables no solo es estadísticamente significativa, sino también relevante desde una perspectiva práctica.

El análisis de residuos tipificados mostró que:

- El depósito Non Refund presenta más cancelaciones de las esperadas.
- El depósito No Deposit presenta menos cancelaciones de las esperadas.
- El depósito Refundable mantiene un comportamiento más equilibrado.

Los resultados evidencian que el tipo de depósito es un factor determinante en la cancelación de reservas. En consecuencia, se recomienda:

- Revisar las condiciones de la modalidad Non Refund, dado su alto nivel de cancelación.
- Incentivar modalidades con menor riesgo de cancelación, como No Deposit o Refundable, a través de promociones o beneficios adicionales.
- Incorporar esta variable en los modelos de predicción de ocupación, de manera que se puedan anticipar cancelaciones y ajustar estrategias de overbooking para optimizar la gestión de ingresos.
- Una adecuada gestión de las políticas de depósito puede reducir pérdidas operativas y mejorar la eficiencia en la planeación hotelera.

Ahora, con la variable ***previous_cancellations***, Se realizó una prueba t de Student para muestras independientes con el fin de comparar el promedio de cancelaciones previas entre dos grupos:

- Clientes que cancelaron su reserva actual.
- Clientes que no cancelaron su reserva actual.

El resultado arrojó un p-value ≈ 4.22e-44, el cual es muy inferior al nivel de significancia ($\alpha = 0.05$). Esto lleva a rechazar la hipótesis nula (H_0) y concluir que existe una diferencia estadísticamente significativa en el número de cancelaciones previas entre los dos grupos.

La diferencia de medias fue de 0.154, con un intervalo de confianza del 95% entre [0.132, 0.175], lo que indica que, en promedio, los clientes que cancelan su reserva actual tienen más cancelaciones previas que aquellos que no lo hacen.

Los resultados evidencian que el historial de cancelaciones pasadas es un predictor relevante del comportamiento de cancelación actual. En términos prácticos:

- Los clientes con más cancelaciones previas tienen mayor probabilidad de volver a cancelar.
- Se recomienda incluir esta variable como un factor de riesgo en los modelos predictivos de cancelación.
- Desde una perspectiva de negocio, podrían establecerse políticas diferenciadas para clientes con historial de cancelaciones (por ejemplo, requerir depósitos más estrictos o restringir beneficios en tarifas flexibles).
- De esta manera, se mejora la gestión del riesgo de cancelaciones y se optimiza la planeación de la capacidad hotelera.

Por otro lado, para la variable ***lead_time*** se aplicó una prueba t de Student para muestras independientes con el fin de comparar el promedio de días de antelación entre los clientes que cancelaron y los que no cancelaron su reserva.

Los resultados muestran un p-value ≈ 0.000, inferior al nivel de significancia ($\alpha = 0.05$). Esto permite rechazar la hipótesis nula (H_0) y concluir que existe una diferencia significativa en el tiempo de antelación entre los dos grupos.

La diferencia de medias encontrada fue de 50.3 días (IC95%: [48.6, 51.9]), indicando que, en promedio, los clientes que cancelan sus reservas lo hacen con aproximadamente 50 días más de anticipación que aquellos que no cancelan.

Este análisis evidencia que reservar con mayor antelación incrementa la probabilidad de cancelación.

A partir de este hallazgo se pueden proponer las siguientes acciones:

- Ajustar políticas de cancelación para reservas hechas con mucha anticipación, aplicando restricciones más estrictas o depósitos más altos.
- Monitorear las reservas tempranas como un factor de riesgo, integrándolas en modelos predictivos de cancelación.
- Ofrecer incentivos para confirmar la estadía (descuentos no reembolsables, upgrades limitados, etc.) a clientes que reservan con mucha anticipación.

En resumen, el lead time constituye un factor clave para predecir cancelaciones y su gestión adecuada puede ayudar a reducir pérdidas por cancelaciones tardías y mejorar la planeación de ocupación hotelera.