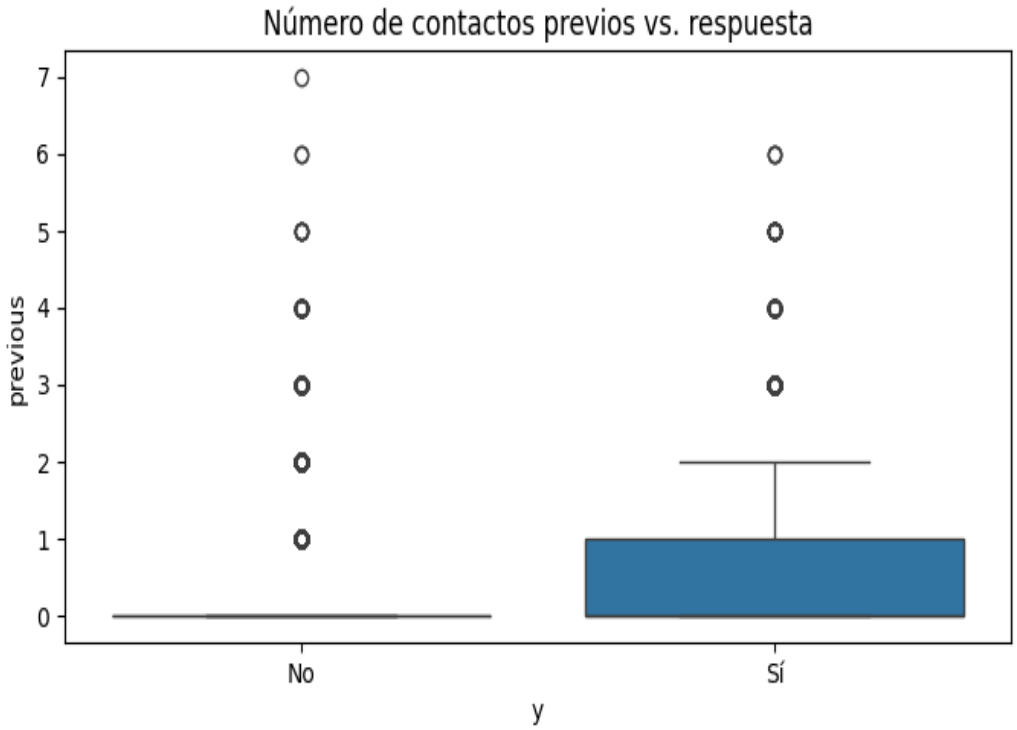
ANÁLISIS DE LOS DATOS OBTENIDOS

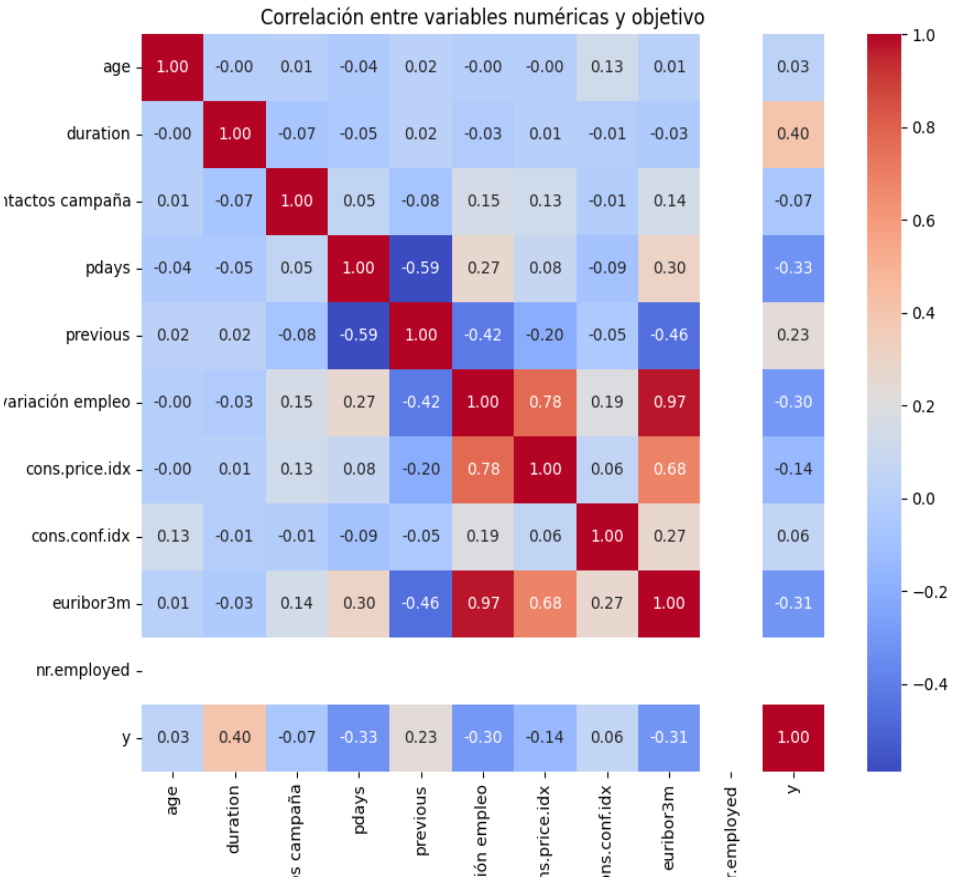
A continuación, realizo un análisis de los gráficos más relevantes, en mi opinión, comenzando por el archivo de **“bank – additional”**



Este gráfico de cajas (boxplot) muestra la relación entre el número de contactos previos realizados con un cliente (“previous”) y la respuesta final del cliente (“y”), categorizada en "Sí" y "No":

- Se observa que los clientes que respondieron "No" suelen tener cero contactos previos, mientras que aquellos que respondieron "Sí" presentan una mayor variabilidad y algunos han tenido varios contactos previos antes de aceptar la oferta.
- La mediana de contactos anteriores para los que respondieron “Sí” es mayor que para los que respondieron “No”.
- Esto sugiere que, en promedio, los clientes que requieren más contactos tienden a aceptar la propuesta, aunque también hay muchos casos en que se logra el “Sí” con pocos contactos.
- Hay presencia de valores atípicos (outliers), especialmente en el grupo "No", lo que indica que algunos clientes han sido contactados muchas veces sin éxito.

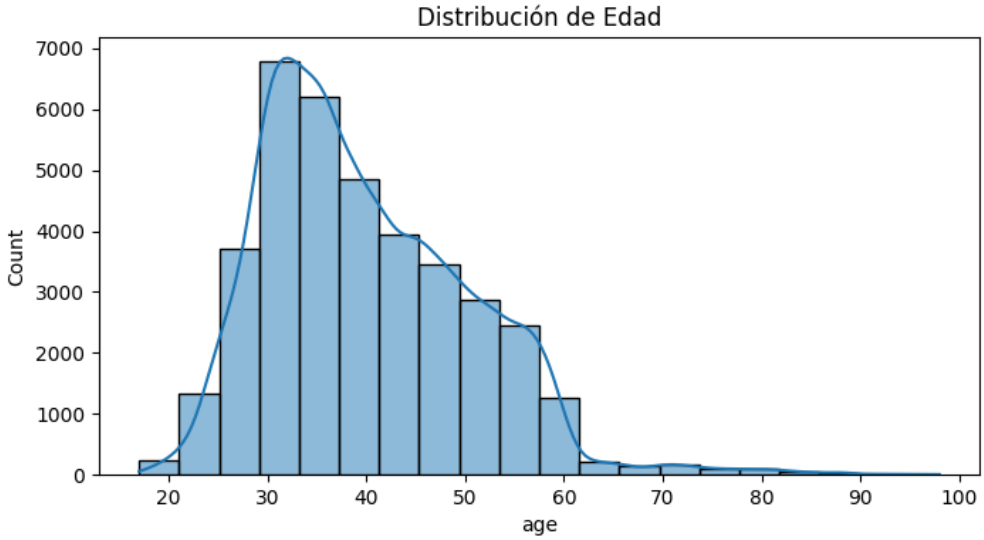
El número de contactos anteriores es un factor relevante a considerar en la estrategia de comunicación, ya que existe una tendencia a obtener mejores respuestas con mayor insistencia, pero también podría ser indicador de saturación y rechazo si se excede el número de intentos.



La matriz de evaluación muestra las relaciones entre variables numéricas del conjunto de datos y el objetivo ("y"):

- La variable con mayor valoración positiva respecto al objetivo es “duración” (0.40), lo que indica que cuanto más largo es el último contacto, mayor probabilidad de respuesta positiva.
- Las variables “pdays” y “previous” muestran correlaciones negativas moderadas con “y”, lo que podría indicar que muchos intentos anteriores o días desde el último contacto disminuyen la probabilidad de éxito.
- Otras variables como “variación empleo”, “cons.price.idx”, “euribor3m” y “nr.empleados” presentan correlaciones moderadas entre sí y algunas correlaciones negativas con el objetivo.
- La edad y los índices de confianza del consumidor tienen correlaciones bajas con el objetivo.

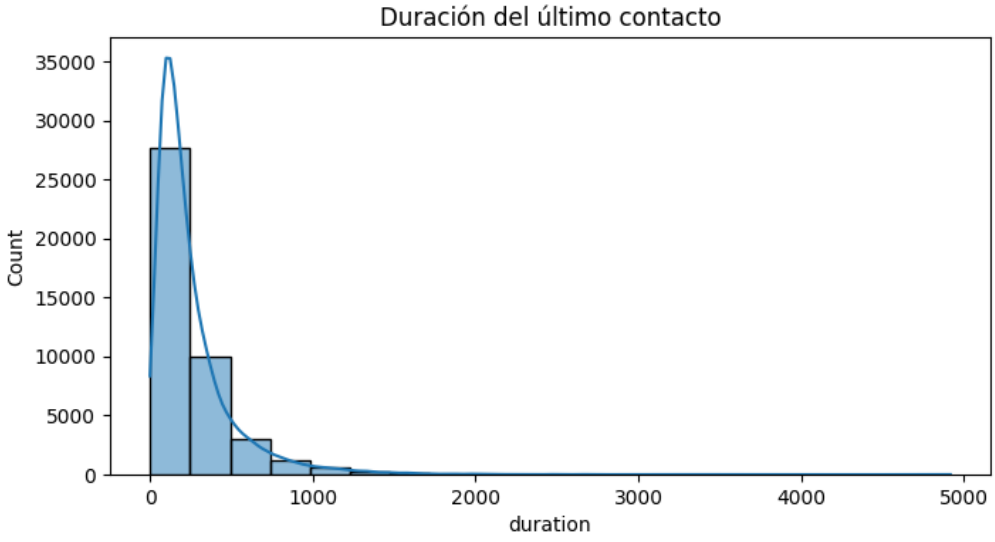
La duración del último contacto es el factor más importante para lograr una respuesta positiva. Las correlaciones entre variables pueden ayudar a identificar redundancias y relaciones para la modelización predictiva.



Este histograma muestra la distribución de edades de los clientes:

- La mayoría de los clientes tienen entre 30 y 40 años, con una clara disminución en frecuencia a medida que aumenta la edad.
- La distribución es asimétrica (sesgo a la derecha), con algunos clientes mayores, pero en menor proporción.
- Hay una pequeña presencia de valores extremos en edades mayores.

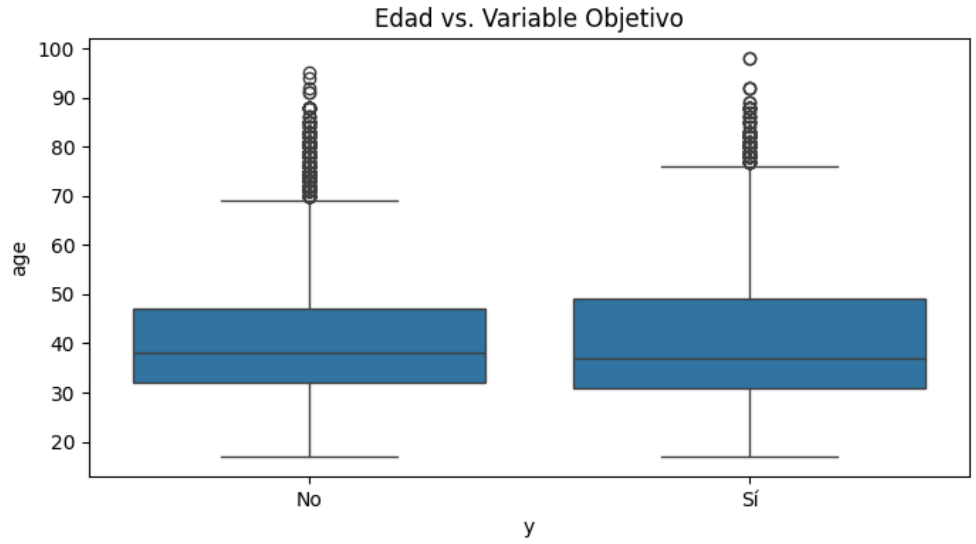
La base de clientes está compuesta principalmente por adultos jóvenes. Las estrategias comerciales podrían adaptarse para aprovechar el segmento dominante y evaluar si los extremos de edad responden de manera diferente.



Este gráfico muestra la distribución de la duración, en segundos, del último contacto realizado con los clientes:

- La mayoría de los contactos tienen una duración corta, por debajo de los 500 segundos.
- Existe una larga cola hacia la derecha, indicando que algunos contactos pueden durar mucho más, pero son casos poco frecuentes.
- La distribución es altamente asimétrica (sesgo a la derecha).

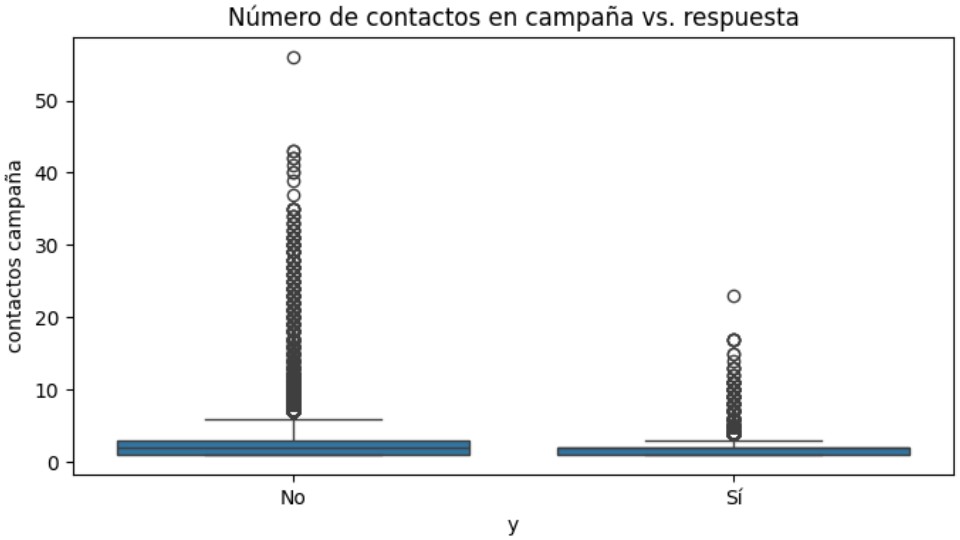
La mayoría de los contactos son breves, lo que puede ser positivo para la eficiencia operativa, pero los contactos más largos tienen mayor relación con respuestas positivas (como se vio en la matriz de compresión). Conviene analizar si optimizar el tiempo de contacto puede mejorar los resultados.



Este gráfico de cajas (boxplot) muestra la distribución de la edad de los clientes en función de la variable objetivo ("y"), que representa la respuesta del cliente a la campaña ("Sí" o "No"):

- Ambas categorías ("Sí" y "No") presentan distribuciones de edad muy similares, con medianas cercanas a los 38-40 años.
- El rango intercuartílico (Q1-Q3) es también comparable para ambos grupos, lo que indica que la mayor parte de los clientes que responden afirmativa o negativamente están en los mismos intervalos de edad.
- Se observan valores atípicos (outliers) en ambos grupos, especialmente en edades superiores a 70 años, lo que indica la presencia de algunos clientes de edad avanzada que han respondido tanto “Sí” como “No”.
- El rango total de edad va aproximadamente de los 18 a los 95 años.
- La dispersión y la simetría del boxplot indican que la edad no presenta una diferencia significativa entre quienes aceptan y quienes rechazan la campaña.

La edad, según este gráfico, no parece ser un factor determinante en la respuesta a la campaña. Tanto los clientes jóvenes como los de mayor edad pueden responder afirmativa o negativamente, y la mediana de edad es prácticamente igual para ambos grupos. Esto sugiere que, en este contexto, segmentar la estrategia comercial únicamente por edad podría no aportar un valor diferencial significativo para mejorar la tasa de éxito.



Este gráfico de cajas (boxplot) muestra la relación entre el número de contactos realizados durante la campaña (“contactos campaña”) y la respuesta final del cliente (“y”), clasificada en "Sí" y "No":

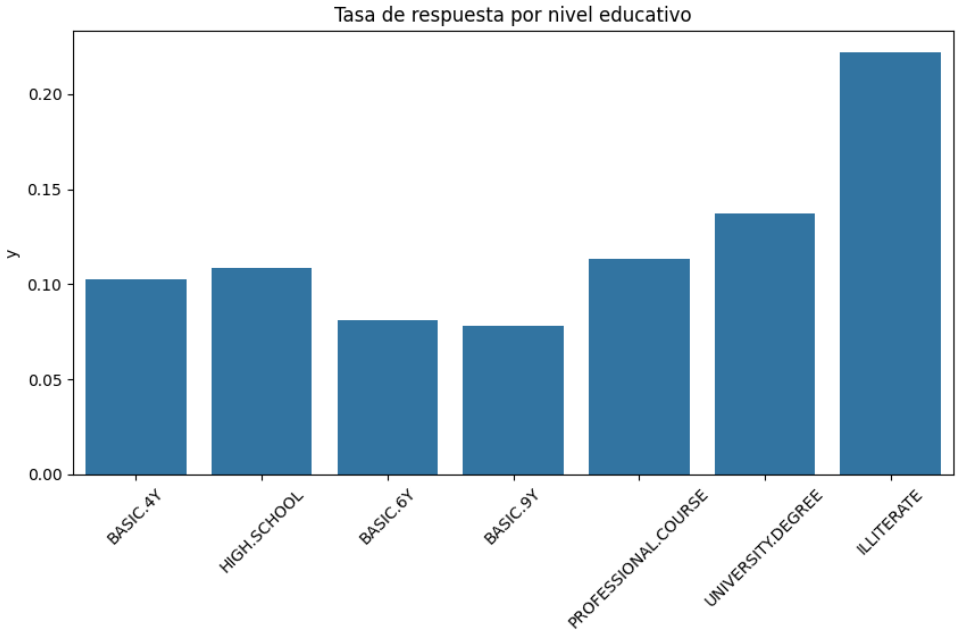
- La mediana del número de contactos es similar en ambos grupos (“Sí” y “No”), con la mayoría de los casos concentrados en valores bajos (1 a 2 contactos).
- Se observa que, para ambos grupos, los valores intercuartílicos son próximos, reflejando que la mayoría de los clientes reciben pocos contactos durante la campaña, independientemente de su respuesta.
- Existen valores atípicos (outliers) bastante notorios, especialmente en el grupo "No", donde algunos clientes han recibido más de 50 contactos sin aceptar la propuesta. En el grupo "Sí" también hay algunos valores atípicos, aunque de menor magnitud.
- La dispersión de los datos es mayor en el grupo “No”, lo que sugiere que insistir excesivamente con algunos clientes no necesariamente mejora la tasa de éxito.

El número de contactos durante la campaña no parece ser un factor decisivo para la respuesta positiva, ya que la mediana es similar entre los que aceptan y los que rechazan. Sin embargo, el gráfico muestra que un exceso de intentos puede estar asociado a una mayor proporción de respuestas negativas, probablemente por saturación o rechazo del cliente. Por lo tanto, se recomienda optimizar el número de contactos y evitar la sobreexposición a los clientes para maximizar la eficiencia y evitar el desgaste.

Este gráfico de barras muestra la tasa de respuesta positiva (“y”) en función de si el cliente tiene o no un préstamo hipotecario:

- La tasa de respuesta es muy similar para ambos grupos, aunque ligeramente superior en el grupo de clientes que sí tienen préstamo hipotecario.
- Los valores de tasa de respuesta rondan el 11% para clientes sin préstamo y el 12% para clientes con préstamo, mostrando una diferencia marginal.
- La variable “préstamo hipotecario” no parece influir de manera significativa en la decisión de los clientes de aceptar la propuesta de la campaña.

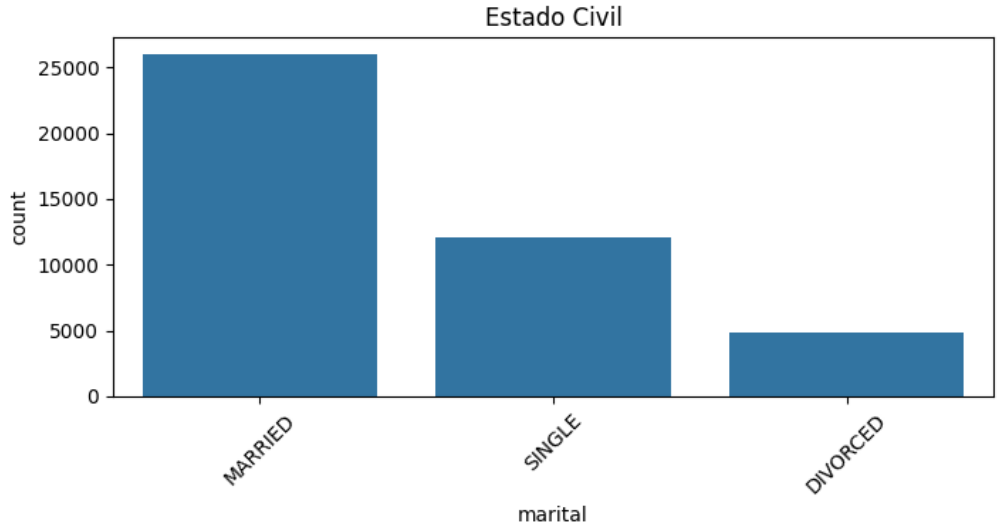
Contratar con un préstamo hipotecario no representa un factor importante para predecir la probabilidad de respuesta positiva a la campaña. La diferencia observada es pequeña y no suficiente para justificar una segmentación exclusiva basada en este criterio. Por tanto, este atributo puede considerarse de bajo impacto en el diseño de estrategias comerciales o en modelos predictivos.



Este gráfico de barras muestra la tasa de respuesta a las llamadas del banco con respecto al nivel educativo de los clientes:

- Los grupos donde más respuestas se producen son los que tienen un alto grado educativo, salvo el grupo de “analfabetos” que superan al resto.
- El grupo de “analfabetos” supera con creces la media de respuestas de los demás grupos.

El mayor grupo con índice de respuesta es el considerado como “analfabetos”, llegando a superar una tasa de respuesta del 20%. Revisando estos datos, se podría pensar que el mayor índice de respuesta estaría entre los grupos con menores estudios, pero sorprende que los siguientes grupos con mayor índice de respuesta sean los que tienen formación universitaria.



Este gráfico de barras muestra la cantidad de préstamos hipotecarios que tienen los clientes en función de su estado civil:

- El grupo con mayor importe de préstamos hipotecarios contratados serían los casados, superando la suma de los dos grupos que completarían la muestra.
- Es representativo que, el grupo de solteros tengan más del doble de préstamos hipotecarios que el grupo de divorciados, siendo este último, el grupo con menor cantidad de préstamos hipotecarios.

Como nos parecería lógico, el grupo de casados es el que más préstamos hipotecarios tienen contratados, ya que, por esa situación familiar podemos entender que contratan préstamos para la casa familiar. Por otro lado, es revelador encontrar al grupo de los solteros como el segundo que más préstamos hipotecarios tiene, superando por el doble, en cantidad, al grupo de los divorciados. Entenderíamos que los solteros optaran por opciones alternativas como el alquiler.

Ahora realizo el análisis de los gráficos más relevantes del archivo **"customer_details".** Para este archivo, se han generada muchos gráficos, ya que son tres años diferentes de datos, por lo que, realizado un análisis general de los datos obtenidos en los diferentes gráficos, que se pueden observar en la carpeta que se incluye para revisión:

Datos principales

Correlación Prácticamente Inexistente entre Variables:

- El hallazgo más significativo y consistente a lo largo de los tres años es la **ausencia casi total de correlación lineal** entre el Ingreso Anual, la actividad en la web y la estructura familiar.
- Los mapas de calor de correlación muestran valores consistentemente cercanos a 0.00, lo que indica que las variables son independientes entre sí.
- Los gráficos de dispersión entre ingresos y visitas web confirman esta falta de relación, mostrando nubes de puntos sin ninguna tendencia discernible.

Distribución del Ingreso Anual:

- El Ingreso Anual se distribuye de forma muy similar en los tres años, abarcando un rango que va aproximadamente de 0 a 180,000.
- Como se ve en los diagramas de caja (box plots), la mediana y el rango de ingresos **no varían** significativamente en función del número de niños, adolescentes en el hogar o el día del mes en que se visita la web.

Composición del Hogar y Actividad Web:

- La distribución de hogares con 0, 1 o 2 niños/adolescentes es relativamente equilibrada en la muestra.
- Las visitas a la web muestran un patrón con picos de actividad a principios y finales de mes. Sin embargo, este comportamiento es independiente del nivel de ingresos o la composición del hogar.

Conclusión

Los datos de 2012, 2013 y 2014 demuestran de manera concluyente que las variables analizadas son **estadísticamente independientes**. Esto significa que, para la población estudiada:

- El nivel de ingresos no predice la cantidad de hijos o la frecuencia de visitas a la web.
- La estructura familiar no está relacionada con el nivel de ingresos.

Cualquier estrategia de negocio que asuma una relación entre estas variables (por ejemplo, dirigir campañas a usuarios de altos ingresos asumiendo que tienen un patrón de navegación específico) no estaría respaldada por este análisis.

Si realizamos un análisis un poco más en profundidad por los tipos de gráfico que obtenemos, podemos obtener las siguientes conclusiones:

1. Inexistencia de Correlación Lineal (Heatmaps y Gráficos de Dispersión)

Los mapas de calor para 2012, 2013 y 2014 muestran de forma consistente coeficientes de correlación de Pearson de **0.00 o 0.01** entre todas las variables. Un coeficiente tan cercano a cero es una fuerte evidencia estadística de que no existe una **relación lineal** entre el ingreso, la composición del hogar y la actividad en la web.

Los gráficos de dispersión refuerzan esta conclusión de manera visual. Muestran nubes de puntos que forman líneas verticales perfectas. Esto indica que para cualquier día del mes en que se visita la web, existe la misma distribución completa de ingresos anuales. Este patrón es tan uniforme que sugiere que los datos podrían ser sintéticos o simulados, ya que en datos del mundo real se esperaría alguna irregularidad.

2. Análisis de las Distribuciones (Histogramas)

- **Ingreso Anual:** La distribución del ingreso es prácticamente **uniforme** en los tres años. Esto significa que hay una cantidad similar de personas en cada tramo de ingresos (por ejemplo, entre 25k-30k, 50k-55k, etc.). Esto es atípico para datos de ingresos reales, que suelen estar sesgados hacia la derecha (más personas con ingresos bajos que altos).
- **Visitas Web al Mes:** La distribución de las visitas web muestra un patrón claro con **picos de actividad en los primeros y últimos días del mes**. Esto podría estar relacionado con ciclos de facturación, pago de salarios o publicación de contenido mensual, aunque sin más contexto es solo una hipótesis.
- **Niños y Adolescentes en el Hogar:** La cantidad de hogares con 0, 1 o 2 niños/adolescentes es casi idéntica. Esto sugiere que la muestra de datos fue estratificada para tener una representación perfectamente equilibrada de estos grupos familiares.

3. Comparación de Grupos (Diagramas de Caja - Box Plots)

- Estos gráficos comparan el Ingreso Anual entre diferentes grupos (según el número de niños, adolescentes o el día de visita a la web).
- En todos los casos, la **línea de la mediana** (el centro de la caja) se mantiene casi perfectamente horizontal a lo largo de las diferentes categorías. Esto demuestra que el ingreso mediano es el mismo, sin importar si una familia tiene 0, 1 o 2 hijos, o si visitan la web el día 5 o el día 25 del mes.
- Además, la altura de las cajas (rango intercuartílico) y la longitud de los bigotes son consistentes, lo que indica que la **dispersión de los ingresos** tampoco cambia entre los grupos.

Implicación de los datos:

La conclusión obtenida en el análisis inicial se confirma: las variables son independientes. Una estrategia de marketing que intente, por ejemplo, dirigirse a usuarios de altos ingresos en días específicos del mes, o que asuma que las familias con más hijos tienen ingresos diferentes, no tendría éxito con esta población, ya que no hay ninguna relación que explotar.