2025

Fundamentos de la Ciencia de Datos

Trabajo Práctico Especial

Grupo 8

Arenaza, Nayla Solange

Astigueta, Ignacio

Lertora, Santiago Tomas

1. Introducción

Este trabajo se centra en uno de los problemas centrales del sector del e-commerce; la predicción de la intención de compra del visitante. En un entorno digital, a diferencia de una tienda física, es difícil medir el "interés" de un cliente. Sin embargo, el comportamiento de navegación deja un rastro de datos (BounceRates, ExitRates, PageValues) que pueden ser analizados.

Comprender los factores que diferencian a una sesión que termina en una transacción (Revenue=True) de una que termina en abandono (Revenue=False) es crucial. Permite a las empresas optimizar sus estrategias de marketing (ej. mostrar ofertas a usuarios indecisos), mejorar el diseño del sitio (ej. simplificar el checkout) y, en última instancia, aumentar la tasa de conversión.

1.1. El Conjunto de Datos: "Online Shoppers Purchasing Intention"

Para este análisis, utilizamos el conjunto de datos online_shoppers_intention.csv. Una investigación online revela que este conjunto fue donado al UCI Machine Learning Repository en 2018.

Contiene 12,330 sesiones de usuarios en un sitio de e-commerce durante un período de un año. El objetivo es predecir si la sesión terminará en una compra (Revenue - "Ganancia"). Las métricas (como BounceRates, ExitRates, PageValues) son derivadas de Google Analytics. De este conjunto de datos, el 84,5% (10,422 sesiones) fueron muestras de la clase negativa (no finalizaron una compra) y el resto, 15,5% (1,908 sesiones), fueron muestras de la clase positiva (si finalizaron con una compra). Esta particularidad (fuerte desbalance de clases) es un punto central del análisis.

El dataset se compone de 18 atributos que caracterizan la sesión: 10 atributos numéricos (Administrative, Administrative_duration, Informational, Informational_duration, ProductRelated, ProductRelated_duration, BounceRates, ExitRates, PageValue, SpecialDay) que miden el comportamiento, y 8 atributos categóricos (OperatingSystems, Browser, Region, TrafficType, VisitorType, Weekend, Month, Revenue) que describen el contexto del visitante.

1.2. Problemas Identificados y Organización del Trabajo

El dataset se compone de 18 atributos. El análisis exploratorio inicial reveló varios problemas claves que requirieron preprocesamiento para asegurar la calidad del análisis. A continuación, se detallan los problemas y las soluciones aplicadas:

- **Tipado Incorrecto de Datos**

- Problema: Variables que son intrínsecamente categóricas, como OperatingSystems, Browser, Region y TrafficType, estaban incorrectamente cargadas como numéricas (int64).
- Solución: Se transformaron estas variables a tipo object (string) para evitar interpretaciones numéricas erróneas y permitir un correcto análisis categórico.

- **Datos Duplicados**
  - Problema: Se identificaron más de 120 filas duplicadas exactas, lo cual es anómalo y podría sesgar el análisis.
  - Solución: Se eliminaron todas las filas duplicadas para hacer un análisis mucho más limpio y garantizar la unicidad de las observaciones.
- **Formato de Tasas y Porcentajes**
  - Problema: Algunas de las variables que representan porcentajes (como BounceRates o ExitRates) estaban expresadas como un valor decimal entre 0 y 1 (por ej.: 0.15).
  - Solución: Para facilitar la interpretación directa, se decidió transformar estas columnas multiplicando sus valores por 100. De esta manera, el valor (por ej.: 0.15) se convierte en 15, representando así el porcentaje real.
- **Granularidad Innecesaria**
  - Problema: La variable SpecialDay (de tipo float) presentaba una granularidad innecesaria, siendo 0.0 en la gran mayoría de los casos.
  - Solución: Se binarizó la variable a un formato booleano (True/False) para simplificar el análisis, donde False representa 0.0 (día no especial) y True cualquier otro valor.
- **Error de inconsistencia**
  - Problema: Observamos que en la variable Month (de tipo object) la mayoria de los meses se encontraban escritos con 3 letras, pero June estaba con 4 letras, como esto podría llevar a un error de consistencia. Para el análisis, "June" y "Jun" (si existiera) serían dos categorías diferentes, lo cual es incorrecto.
  - Solución: Se decidió reemplazar “June” por “Jun” para que los datos queden consistentes.
- **Inconsistencia de categoría**
  - Problema: Se detectó que la variable VisitorType contenía 85 instancias (representando solo el 0.69% del dataset) de la categoría "Other". Esta categoría no se encuentra en la descripción de este atributo, tal que, en el mismo, los valores solo podían ser “Returning_Visitor” o “New_Visitor”, por lo tanto, “Other” no debería ser parte de los resultados.
  - Solución: eliminar las 85 filas correspondientes a esta categoría.
- **Presencia de Outliers**

- Problema: Varias variables numéricas, principalmente las relacionadas con tiempos de navegación y número de páginas visitadas, presentaban valores máximos extremadamente alejados del comportamiento típico de los usuarios (por ejemplo, sesiones de más de 17 horas o navegación de cientos de páginas en una sola visita). Estos valores anómalos distorsionaban las medidas estadísticas y podían sesgar los análisis posteriores.
- Solución: Se aplicó un método de *capping* mediante el percentil 99, reemplazando únicamente los valores por encima de ese umbral por el propio límite. Esto permitió corregir valores irrealistas sin eliminar registros completos y mejorar la estabilidad del dataset para etapas posteriores del análisis.

Finalmente, este informe se organiza:

- **Sección 2 (Análisis Exploratorio de los datos):** Se presenta la caracterización detallada de los atributos, sus distribuciones visuales, análisis de outliers y los hallazgos iniciales. Como también, se justifican y ejecutan las transformaciones y eliminaciones de datos.
  - **Sección 2.1 (Análisis de cada una de las variables)**: Se describen las variables del dataset, analizando sus distribuciones, tipos y comportamientos principales, para identificar patrones, sesgos y posibles problemas en los datos.
  - **Sección 2.2 (Procesamiento de Outliers)**: Se corrigen valores extremos en variables mediante *capping* con el percentil 99.
- **Sección 3 (Hipótesis planteadas y resolución):** Se plantean y validan las 6 hipótesis.
- **Sección 4 (Conclusión):** Se resumen los hallazgos clave del trabajo y se reflexiona sobre los resultados.

2. Análisis exploratorio de los datos

Para el análisis exploratorio de los datos comenzamos viendo cuantos datos tenemos y de que tipo son:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12330 entries, 0 to 12329
Data columns (total 18 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Administrative           12330 non-null  int64
 1   Administrative_Duration  12330 non-null  float64
 2   Informational            12330 non-null  int64
 3   Informational_Duration   12330 non-null  float64
 4   ProductRelated           12330 non-null  int64
 5   ProductRelated_Duration  12330 non-null  float64
 6   BounceRates              12330 non-null  float64
 7   ExitRates                12330 non-null  float64
 8   PageValues               12330 non-null  float64
 9   SpecialDay               12330 non-null  float64
 10  Month                    12330 non-null  object
 11  OperatingSystems         12330 non-null  int64
 12  Browser                  12330 non-null  int64
 13  Region                   12330 non-null  int64
 14  TrafficType              12330 non-null  int64
 15  VisitorType              12330 non-null  object
 16  Weekend                  12330 non-null  bool
 17  Revenue                  12330 non-null  bool
dtypes: bool(2), float64(7), int64(7), object(2)
memory usage: 1.5+ MB
```

***Tabla 1**: Nos da la información sobre nuestro DataFrame, la cantidad de registros, el número de columnas, cantidad de valores nulos en cada columna y el tipo de dato asociado a cada una de las columnas*

En primer lugar, podemos observar que cuenta con 18 columnas y 12330 registros, con 12330 registros no nulos en todas las columnas. El análisis de la columna Dtype revela una discrepancia. Mientras que los atributos numéricos (float64, int64) y algunas categóricas (object, bool) son correctos, existe un grupo de variables (OperatingSystems, Browser, Region y TrafficType) que están cargadas como int64. Conceptualmente, estas variables son categóricas (IDs que representan grupos) y tratarlas como números enteros podría llevar a interpretaciones erróneas. Para confirmar esta discrepancia y, a la vez, analizar el comportamiento, la distribución y la presencia de outliers en las variables numéricas, se generó la siguiente tabla de estadísticas descriptivas:

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	OperatingSystems	Browser	Region	TrafficType
count	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000
mean	2.315166	80.818611	0.503569	34.472398	31.731468	1194.746220	0.022191	0.043073	5.889258	0.061427	2.124006	2.357097	3.147364	4.069586
std	3.321784	176.779107	1.270156	140.749294	44.475503	1913.669288	0.048488	0.048597	18.568437	0.198917	0.911325	1.717277	2.401591	4.025169
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000
25%	0.000000	0.000000	0.000000	0.000000	7.000000	184.137500	0.000000	0.014286	0.000000	0.000000	2.000000	2.000000	1.000000	2.000000
50%	1.000000	7.500000	0.000000	0.000000	18.000000	598.936905	0.003112	0.025156	0.000000	0.000000	2.000000	2.000000	3.000000	2.000000
75%	4.000000	93.256250	0.000000	0.000000	38.000000	1464.157214	0.016813	0.050000	0.000000	0.000000	3.000000	2.000000	4.000000	4.000000
max	27.000000	3398.750000	24.000000	2549.375000	705.000000	63973.522230	0.200000	0.200000	361.763742	1.000000	8.000000	13.000000	9.000000	20.000000

***Tabla 2**: Serie de estadísticas descriptivas de todas las columnas, ya sea el desvió estándar (std), valor mínimo de esa columna (min), etc.*

Como podemos observar la tabla confirma la discrepancia, tal que se calculan estadísticas (mean, std, min, max, etc.) para OperatingSystems, Browser, Region y TrafficType. Estos cálculos carecen de sentido conceptual y demuestran la necesidad de transformar estas columnas a tipo object. Además, Los valores

en estas columnas (ej. Region = 1, Region = 2) no son cantidades matemáticas, sino etiquetas que representan grupos discretos entre los cuales no existe una relación de magnitud. Por lo tanto, se procedió a convertir estas variables del tipo int64 a object para su correcto tratamiento.

En segundo lugar, notamos que las variables que describen porcentajes (como BounceRates y ExitRates) estaban expresadas en formato de valores entre 0 y 1. Para facilitar la interpretación en los análisis y visualizaciones, se transformaron estas columnas multiplicando todos sus valores por 100. De esta manera, podemos trabajar con un formato de porcentaje más intuitivo (por ejemplo, 0.15 se convierte en 15).

En tercer lugar, encontramos que existían filas con datos duplicados. Se identificaron más de 120 registros completamente repetidos, lo cual es anómalo y puede introducir sesgos al mantener observaciones idénticas en el análisis. Para asegurar la consistencia del dataset y preservar la unicidad de las sesiones, se eliminaron todas las filas duplicadas antes de continuar con el procesamiento.

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay
count	12205.000000	12205.000000	12205.000000	12205.000000	12205.000000	12205.000000	12205.000000	12205.000000	12205.000000	12205.000000
mean	2.338878	81.646331	0.508726	34.825454	32.045637	1206.982457	2.037032	4.146560	5.949574	0.061942
std	3.330436	177.491845	1.275617	141.424807	44.593649	1919.601400	4.525544	4.616270	18.653671	0.199666
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	8.000000	193.000000	0.000000	1.423097	0.000000	0.000000
50%	1.000000	9.000000	0.000000	0.000000	18.000000	608.942857	0.289855	2.500000	0.000000	0.000000
75%	4.000000	94.700000	0.000000	0.000000	38.000000	1477.154762	1.666667	4.852941	0.000000	0.000000
max	27.000000	3398.750000	24.000000	2549.375000	705.000000	63973.522230	20.000000	20.000000	361.763742	1.000000

***Tabla 3**: Serie de estadísticas descriptivas de todas las columnas, ya sea el desvió estándar (std), valor mínimo de esa columna (min), etc. Luego, de las transformaciones.*

2.1 Análisis de cada una de las variables

La variable **Administrative** describe el número de páginas de tipo "administrativo" que visito el usuario en esa sesión. El análisis de esta variable revela una distribución fuertemente asimétrica (sesgo a la derecha):

- El histograma muestra un fuerte sesgo positivo, tal que la gran mayoría de las observaciones se concentran en valores muy bajos (cerca del 0), con una "cola" larga que se extiende hacia valores más altos. Esto se confirma al observar que la media (2.32) es significativamente mayor que la mediana (1.0).
- Esta concentración en valores bajos se vuelve aún más evidente al observar los cuartiles: el primer cuartil (25%) es 0.0 y la mediana (50%) es 1.0. Esto es un dato muy potente, ya que nos dice que al menos el 25% de las sesiones no tuvieron *ninguna* visita administrativa, y que la mitad de todas las sesiones (el 50%) tuvieron 1 visita o menos.

- El boxplot muestra que la "caja" esta comprimida en un rango pequeño (de 0 a 4). Mas importante aún, visualiza la "cola larga" del histograma como una gran cantidad de valores atípicos (outliers), que son todos los puntos que se disparan hasta el valor máximo de 27.
- La desviación estándar (3.32) es alta en comparación con la mediana (1.0) e incluso mayor que la propia media (2.32). Esto indica que, aunque la mayoría de los datos están agrupados, existe una gran dispersión general, causada precisamente por esos valores atípicos.

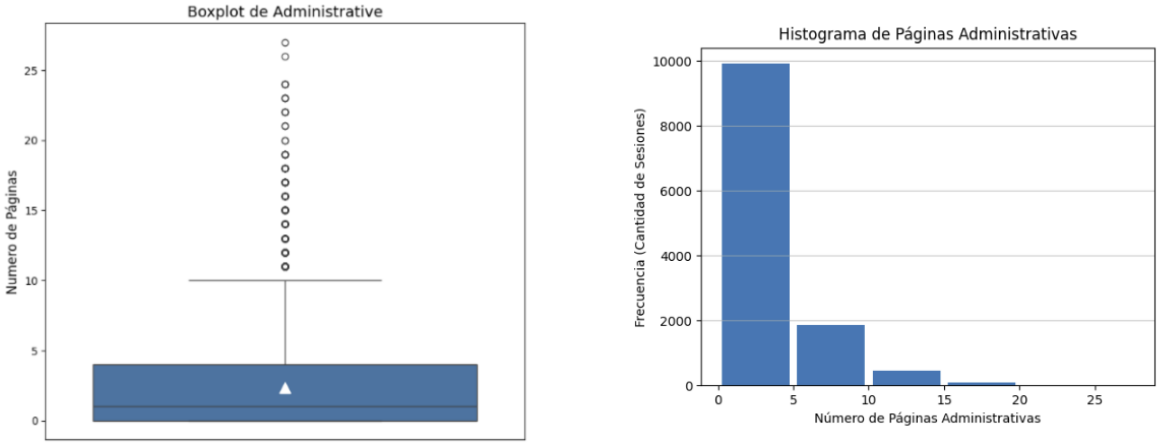

***Figura 1 y 2**: Análisis estadístico y visual de la variable Administrative**.** Se observa un fuerte sesgo positivo y numerosos valores atípicos.*

La variable **Administrative_duration** describe el tiempo total (en segundos) que el usuario paso sumando todas las páginas administrativas que visito. El análisis de la variable revela una distribución fuertemente asimétrica (sesgo a la derecha):

- El histograma está casi completamente colapsado en el valor 0. Muestra una barra masiva en cero, con una "cola" larguísima y casi invisible que se extiende hacia valores muy altos.
- La diferencia entre la mediana (0.0) y la media (80.82) son muy distintas, tal que la mediana es la medida más representativa, indicando que al menos el 50% de todas las sesiones pasaron 0 segundos en páginas administrativas. Mientras que en la media de 80.82 segundos es un valor completamente distorsionado por los valores atípicos y no representan al usuario típico.
- El Primer Cuartil (25%) como la Mediana (50%) son 0.0, mientras que el Tercer Cuartil (75%) es de solo 12.0 segundos, tal que esto nos dice que el 75% de todas las sesiones de usuarios duraron 12 segundos o menos en páginas administrativas.
- El diagrama de boxplot se puede ver que la “caja” es extremadamente delgada, comprimida entre 0 y 12. Además, en este diagrama está dominado por una enorme cantidad de valores atípicos, que están por encima del bigote (encima de 12) y se extienden hasta el máximo de 3398.75, es decir casi 57 minutos.

- La desviación estándar (176.78) lo cual podemos rescatar que es más del doble de la media, por lo tanto, concluimos que está fuertemente influenciada por los valores atípicos; su valor es tan alto únicamente porque esos pocos puntos están extremadamente lejos de la masiva concentración en 0.

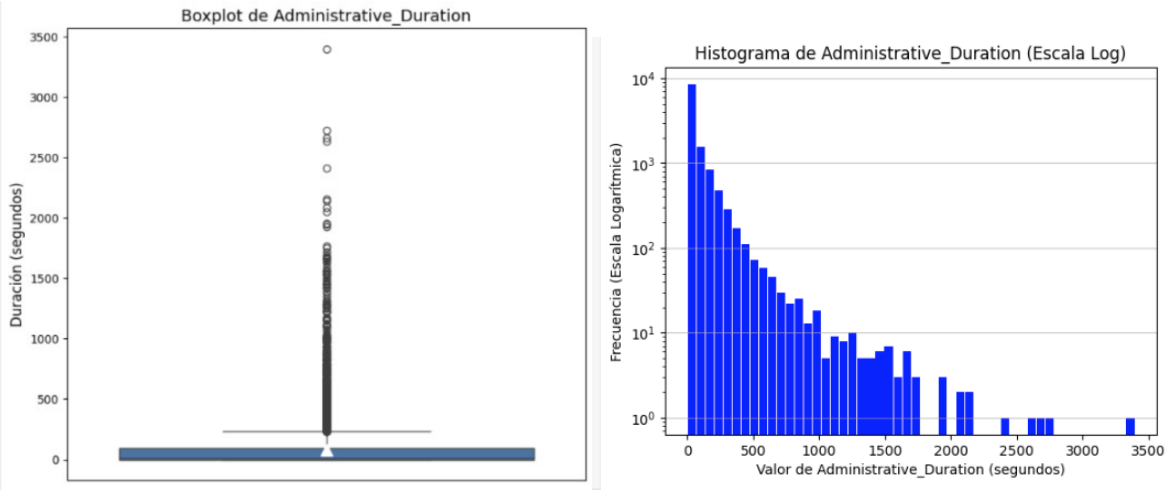

Figura 3 y 4: *Análisis estadístico y visual de la variable Administrative_duration. Se observa un fuerte sesgo positivo y numerosos valores atípicos.*

La variable **Informational** describe el número de páginas de tipo "informativo" que visito el usuario en esa sesión. La distribución de esta variable es un caso extremo de asimetría positiva (sesgo a la derecha):

- El histograma está completamente colapsado en una única barra en el valor 0. Esto indica que la inmensa mayoría de las observaciones tienen este valor, y cualquier otro valor es extremadamente infrecuente.
- La diferencia entre la media (0.50) y la mediana (0.0) son muy distintitas, tal que la mediana es la única mediana representativa de la tendencia central, indicando que la mitad de los usuarios (o más) no visito ninguna página informativa. Mientras que la media está completamente distorsionada por los pocos valores no nulos (los outliers) y no describe al usuario típico.
- El Primer Cuartil (25%), la Mediana (50%) y el Tercer Cuartil (75%) son todos 0.0, lo cual esto significa que al menos el 75% de todas las sesiones de usuarios tuvieron 0 visitas a páginas de tipo “Informational”.
- En el diagrama de Boxplot se puede ver que la “caja” representa el 50% central de los datos, no existe, ya que es solo una línea en el 0. Los otros puntos que podemos observar en el diagrama son valores atípicos (outlier).

- La desviación estándar es relativamente alta porque está fuertemente influenciada por los valores atípicos. Aunque la gran mayoría de los datos se concentra en 0, la existencia de valores tan lejanos como 12 distorsiona la métrica, resultando en un promedio de dispersión de 1.27.

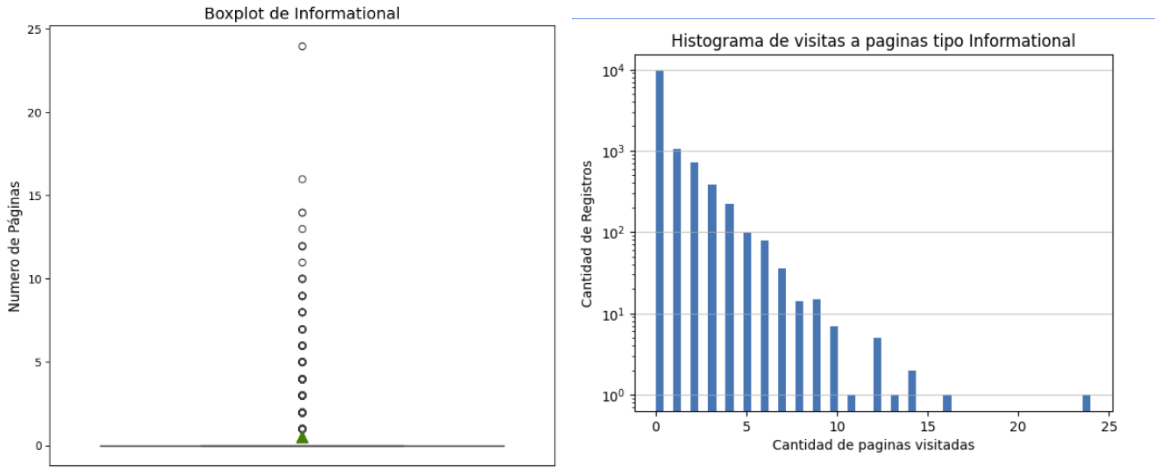

***Figura 5 y 6**: Análisis estadístico y visual de la variable Administrative_duration. Se observa un fuerte sesgo positivo y numerosos valores atípicos.*

La variable **Informationa_duration** describe el tiempo total (en segundos) que el usuario paso sumando todas las páginas informativas que visito. El análisis de la variable revela una distribución fuertemente asimétrica (sesgo a la derecha):

- El histograma está completamente colapsado en el valor 0. Muestra una única barra masiva en 0, y una "cola" imperceptible que se extiende hasta el valor máximo.
- Hay una gran discrepancia entre la mediana (0.0) y la media (34.42), tal que esta última está completamente distorsionada por los valores atípicos y no describe al usuario típico. Además, la mediana (0.0) es la unca medida de tendencia central que describe la realidad de los datos, la cual representa que la mitad de las sesiones (o más) tuvieron 0 segundos de duración.
- El Primer Cuartil (25%), la Mediana (50%) y el Tercer Cuartil (75%) son todos 0.0, esto significa que al menos el 75% de todas las sesiones de usuarios pasaron 0 segundos en páginas de tipo “Informational”.
- El diagrama de boxplot se puede ver que la “caja” no existe; es solo una línea en el 0, confirmando que el Rango Intercuartílico (IQR) es 0, por lo que cualquier valor no nulo se considera estadísticamente valor atípico (outlier). El diagrama muestra una densa nube de estos valores atípicos, que se extienden hasta el máximo de 2549.37 segundos (casi 42.5 minutos).

- La desviación estándar (140.74) es altísima (más de 4 veces la media). Este valor está fuertemente inflado por los outliers; no representa una dispersión general, sino el efecto desproporcionado de esos pocos puntos que están extremadamente lejos de la masiva concentración en 0.

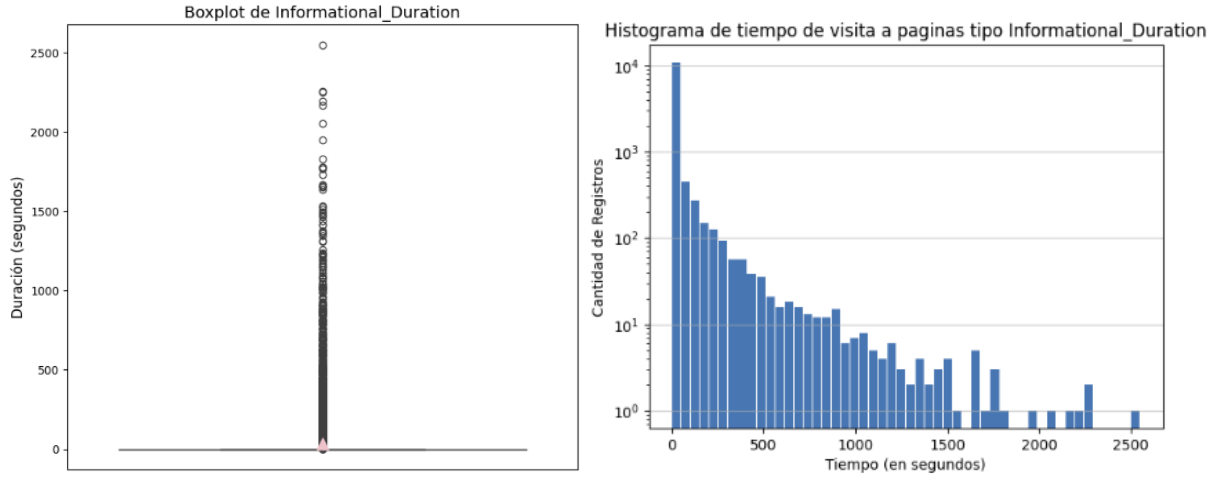

***Figura 7 y 8**: Análisis estadístico y visual de la variable informationa_duration. Se observa un fuerte sesgo positivo y numerosos valores atípicos.*

La variable **ProductRelated** describe el número de páginas "relacionadas a productos" que visito el usuario en esa sesión. El análisis de la variable revela una distribución fuertemente asimétrica (sesgo a la derecha):

- El histograma muestra un pico claro en valores bajos (el "grueso" de los datos está entre 0 y 50), seguido de una "cola" muy larga que se extiende hacia la derecha.
- La media (31.73) es significativamente más alta que la mediana (18.0). Esto es un efecto clásico de los valores extremos (outliers) en la cola, que "tiran" del promedio hacia arriba. La mediana (18 páginas) es la medida más representativa del "comportamiento típico" de un usuario en esta variable.
- El Primer Cuartil (25%) es 7.0. Esto nos dice que el 25% de los usuarios visitó al menos 7 páginas de productos. Además, el Rango Intercuartílico (IQR) es 31.0, esto significa que el 50% central de todos los usuarios visitó entre 7 y 38 páginas de productos.
- En el diagrama de boxplot se puede ver que la “caja” está definida entre el 7 y 38, pero que está en la parte baja de la escala general. Además, en el diagrama se puede ver que está dominado por una masiva cantidad de valores atípicos (outliers), que son todos los puntos que se disparan por encima del bigote superior (aproximado 85) y se extienden hasta el máximo de 705.

- La desviación estándar (44.47) es extremadamente alta (¡incluso más alta que la media!). Este valor está fuertemente inflado por los outliers. No describe la dispersión del 50% central, sino el efecto desproporcionado de esos pocos usuarios que visitaron cientos de páginas.

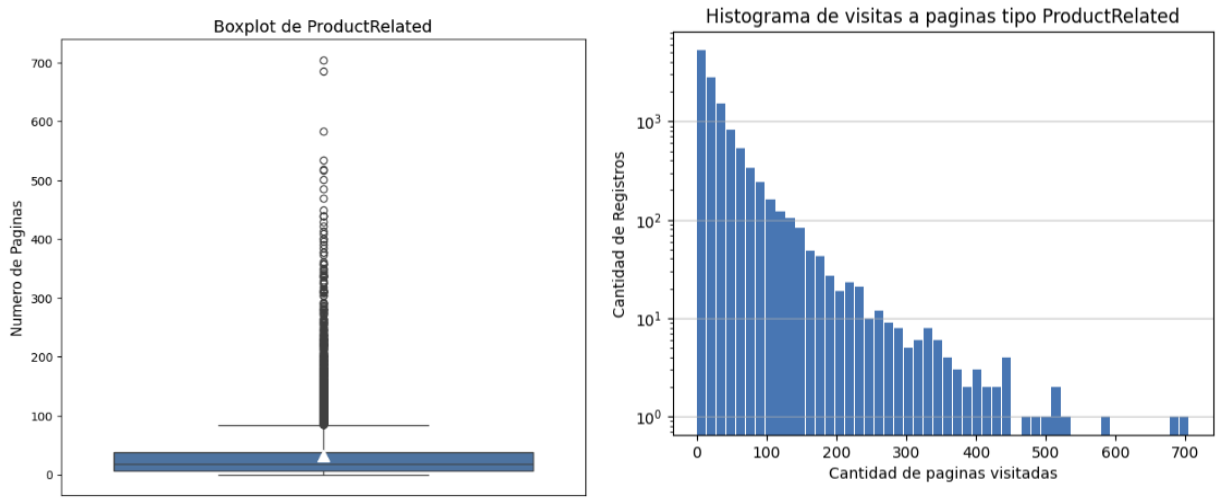

Figura 9 y 10: *Análisis estadístico y visual de la variable ProductRelated. Se observa un fuerte sesgo positivo y numerosos valores atípicos.*

La variable **ProductRelated_duration** describe el tiempo total (en segundos) que el usuario paso sumando todas las páginas de productos. El análisis de la variable revela una distribución fuertemente asimétrica (sesgo a la derecha):

- El histograma muestra un pico claro en valores bajos (el "grueso" de los datos está entre 0 y aproximadamente 2500 segundos), seguido de una "cola" extremadamente larga que se extiende hacia la derecha.
- La media (1194.75) es casi exactamente el doble que la mediana (598.94). Este es un signo clásico de que la media está fuertemente distorsionada (inflada) por los valores extremos en la cola. Además, la mediana (598.94 s aproximadamente 10 minutos) es la medida más representativa del "tiempo típico" que pasa un usuario.
- El Primer Cuartil (25%) es 184.14 (aprox. 3 minutos). Tal que, el Rango Intercuartílico (IQR) es 1280.02. Esto significa que el 50% central de todos los usuarios (el "cuerpo" principal de los datos) pasó entre 184.14 y 1464.16 segundos (aproximadamente 3 a 24 minutos).
- En el diagrama de Boxplot se puede ver que la “caja” está definida en el IQR, pero que está en la parte baja de la escala general. Además, se puede observar que el grafico esta domino por una masiva y densa nube de valores atípicos (outliers) son todos los puntos que se disparan por

encima del bigote superior (aprox. 3384 s) y se extienden hasta el máximo de 63973.52 segundos aproximadamente 17 horas.

- La desviación estándar (1913.67) es grande (incluso mayor que la media). Este valor está fuertemente inflado por los outliers. No describe la dispersión del 50% central, sino el efecto desproporcionado de esos pocos usuarios que pasaron una cantidad de tiempo extrema en el sitio.

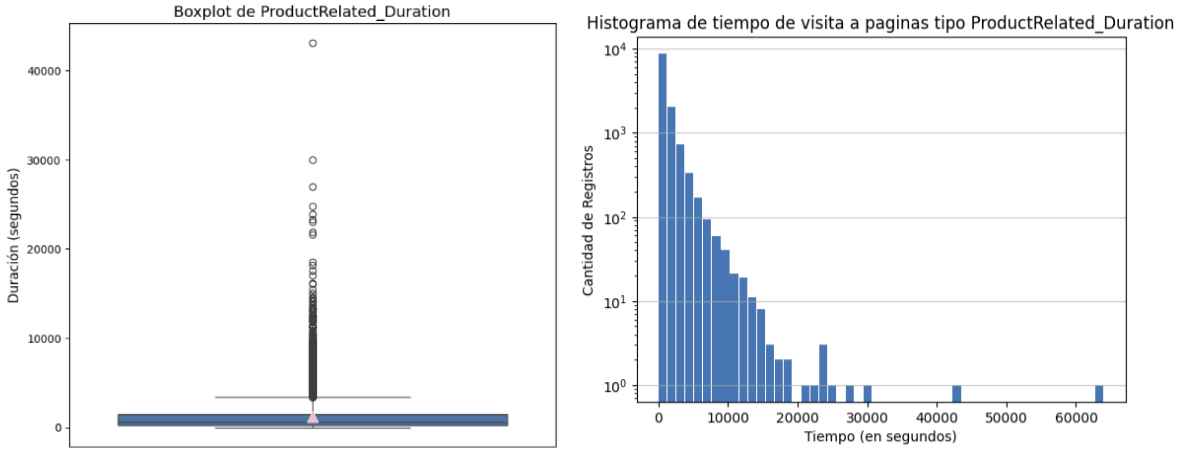

Figura 11 y 12: *Análisis estadístico y visual de la variable ProductRelated_duration. Se observa un fuerte sesgo positivo y numerosos valores atípicos.*

La variable **BounceRates** se refiere al porcentaje de visitantes que entran al sitio y luego la abandonan sin generar ninguna solicitud al servidor de análisis durante esa sesión. El análisis de la variable revela una distribución fuertemente asimétrica (sesgo a la derecha):

- El histograma muestra una masiva concentración de datos en el valor 0 o muy cerca de él. Hay una "cola" larga, aunque no tan extrema como en casos anteriores, que se extiende hacia la derecha.
- La diferencia entre la mediana (0.3) y la media (2.22) es el indicador de esta asimetría, tal que la media es mas de 7 veces mayor que la mediana esto nos indica que la media está fuertemente distorsionada por los valores más altos en la cola. Además, la mediana es mucho más representativa del comportamiento central indicando que la mitad de las sesiones tuvieron un porcentaje de rebote de 0.3% o menos.
- El Primer Cuartil (25%) es 0.0. Esto significa que al menos el 25% de las sesiones no tuvieron ningún rebote. El Tercer Cuartil (75%) es 1.6. Esto nos dice que el 75% de todas las sesiones tuvieron una tasa de rebote inferior al 1.6%.

- En el diagrama boxplot se puede ver que la “caja” está bastante comprimida en un rango pequeño (0 a aproximadamente 1.6). Además, en el diagrama se muestra una gran cantidad de valores atípicos, que están por encima del bigote superior (aproximado 4.0) y se extiende hasta el máximo de 20.0.
- La desviación estándar (4.85) es muy alta (más del doble de la media), este valor está fuertemente inflado por los outliers; la dispersión real para la mayoría de los datos es baja, pero los valores extremos distorsionan esta métrica.

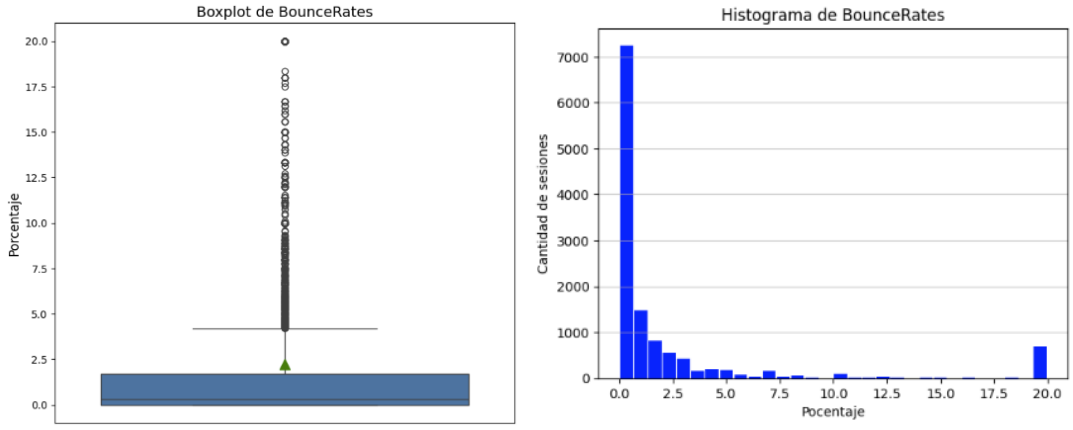

***Figura 13 y 14**: Análisis estadístico y visual de la variable BounceRates. Se observa un fuerte sesgo positivo y numerosos valores atípicos.*

La variable **ExitRates** se calcula como el porcentaje de páginas vistas que fueron las últimas de la sesión, considerando todas las páginas vistas. A diferencia de muchas de las variables anteriores que estaban colapsadas en cero, ExitRates tiene una distribución más "completa", aunque sigue siendo fuertemente asimétrica (sesgo a la derecha):

- El histograma muestra que la mayoría de los datos se agrupan en valores bajos (el pico está alrededor de 1.0 a 3.0), pero existe una "cola" larga que se extiende hacia el valor máximo de 20.
- La media (4.31) es significativamente mayor que la mediana (2.5). Esto ocurre porque los valores altos en la cola "tiran" del promedio hacia la derecha. La mediana (2.5) es una medida mucho más representativa de los porcentajes de salida "típica" de una sesión.
- El Primer Cuartil (25%) es 1.45, a diferencia de BounceRates, esta variable no está dominada por ceros. Además, El 75% de las sesiones tiene una tasa de salida superior a 0.
- El Rango Intercuartílico (IQR) es 3.25, tal que la "caja" del boxplot nos dice que el 50% central de todas las sesiones tiene una tasa de salida que se encuentra en el rango de 1.45% a 4.7%. Además, el diagrama nos muestra que hay una gran cantidad de valores atípicos, que son todos

los puntos que se disparan por encima del bigote superior (aproximadamente desde 10.2) y se extiende hasta el máximo de 20.0.

- La desviación estándar (4.85) es alta (más alta que la media). Este valor está fuertemente inflado por los outliers, indicando que, aunque el “cuerpo” de los datos esta compacto, la dispersión general es grande debido a la “cola”.

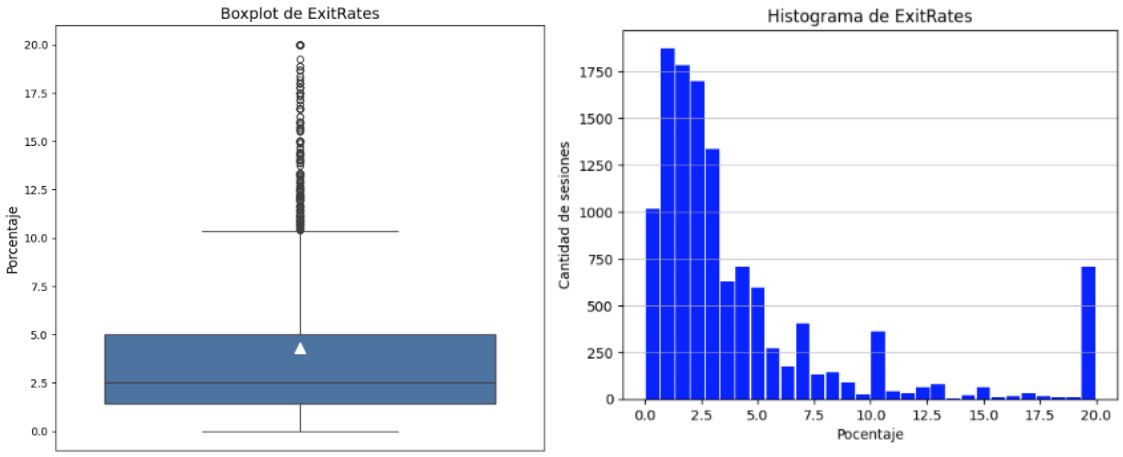

***Figura 15 y 16**: Análisis estadístico y visual de la variable ExitRates. Se observa un fuerte sesgo positivo y numerosos valores atípicos.*

La variable **PageValue** se representa el valor promedio de una página web que un usuario visitó antes de completar una transacción de comercio electrónico. El análisis de la variable revela una distribución fuertemente asimétrica (sesgo a la derecha):

- El histograma está completamente colapsado en el valor 0. Muestra una única barra masiva en 0, lo que indica que la inmensa mayoría de las sesiones no tienen PageValue (probablemente porque no completan una transacción).
- La mediana (0.0) es la única medida que describe la realidad de la gran mayoría de los datos. Mientras que la media (5.88) es una estadística que existe únicamente debido a los valores atípicos. Es un valor completamente engañoso si se usa para describir al usuario "típico".
- El Primer Cuartil (25%), la Mediana (50%) y el Tercer Cuartil (75%) son todos 0.0. Esto significa que al menos el 75% de todas las sesiones de usuario tienen un PageValue de 0.
- En el diagrama de Boxplot se puede ver que la “caja” no existe; es solo una línea en el 0, confirmando que el Rango Intercuartílico (IQR) es 0, tal que cualquier valor no nulo (mayor a 0) se considera estadísticamente un valor atípico (outlier). El gráfico muestra una densa nube de estos outliers, que se extienden hasta el máximo de 361.76.

- La desviación estándar (18.57) es grande (más de 3 veces la media). Este valor está fuertemente inflado por los outliers. No representa una dispersión general, sino el efecto desproporcionado de esos pocos puntos que están extremadamente lejos de la masiva concentración en 0.

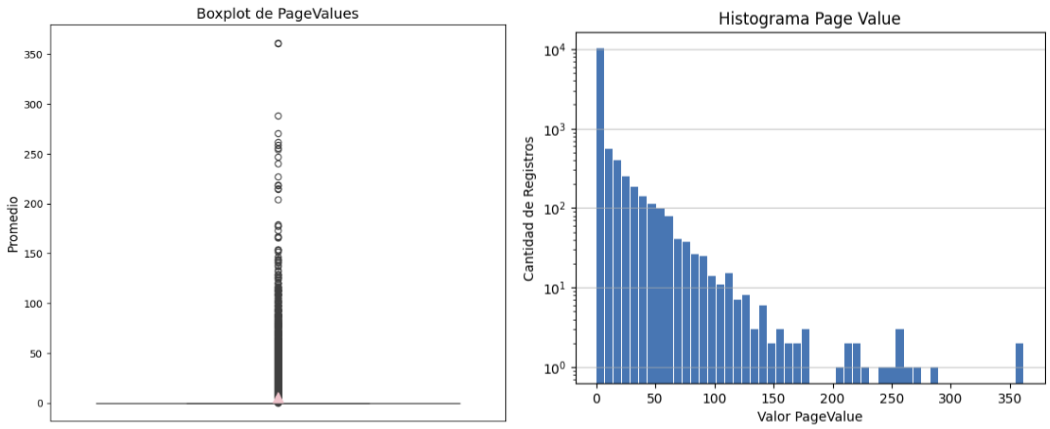

***Figura 17 y 18**: Análisis estadístico y visual de la variable PageValue. Se observa un fuerte sesgo positivo y numerosos valores atípicos.*

La variable **SpecialDay** mide la cercanía de la fecha de la visita a un día festivo relevante para el e-commerce (como Día de San Valentín, Día de la Madre, etc.). La distribución de SpecialDay es un caso de **asimetría positiva extrema:**

- El histograma está casi completamente colapsado en el valor 0.0. Muestra una única barra masiva en 0, lo que indica que la inmensa mayoría de las sesiones no ocurrieron cerca de un día festivo. Las otras barras (en 0.2, 0.4, etc.) son tan pequeñas que son casi invisibles.
- La mediana (0.0) es la única medida que describe la realidad de la gran mayoría de los datos. Mientras que la media (0.062) es un valor muy bajo, pero aun así es engañoso, ya que solo existe debido a los pocos valores no nulos.
- El Primer Cuartil (25%), la Mediana (50%) y el Tercer Cuartil (75%) son todos 0.0. Esto significa que al menos el 75% de todas las sesiones de usuario tienen un valor de 0.0 para SpecialDay. Dado que el IQR es 0, cualquier valor no nulo (0.2, 0.4, 0.6, 0.8, 1.0) se considera estadísticamente un valor atípico (outlier).

Basándonos en el análisis anterior, reveló un problema de granularidad innecesaria. Aunque la variable es de tipo float (con valores de 0.0 a 1.0), el análisis estadístico (Q1=0.0, Q3=0.0) confirma que al menos el 75% de las observaciones son 0.0. Los valores restantes (0.2, 0.4, etc.) son tan infrecuentes que añaden complejidad sin valor. Para simplificar el rasgo y capturar su significado real (si es o no un día especial), se

decidió binarizar la variable. Se transformó la columna a un tipo booleano, donde 0.0 se asignó a False (día no especial) y cualquier valor > 0.0 se asignó a True (día especial).

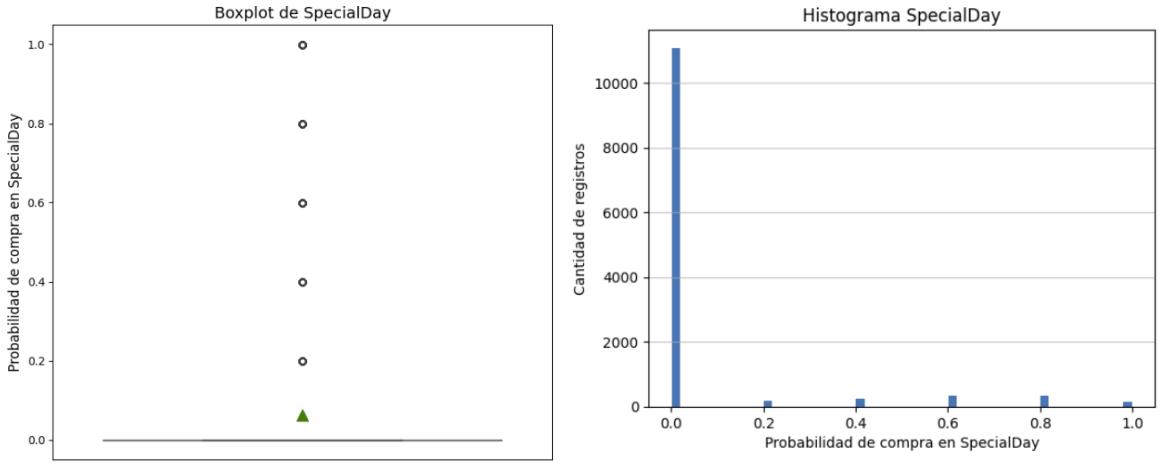

Figura 19 y 20: *Análisis estadístico y visual de la variable SpecialDay. Se observa una asimetría positiva extrema.*

Tras analizar la variable SpecialDay, observamos que su transformación a formato booleano reducía la variabilidad original (valores como 0.2, 0.4 o 0.6), provocando que distintas observaciones pasaran a ser idénticas. Esto generó un pequeño conjunto de filas duplicadas que no existían en el dataset original. Dado que estos duplicados eran producto exclusivo de la transformación y no aportaban información nueva, y además correspondían únicamente a 5 filas. Entonces, se decidió eliminarlos para mantener la coherencia y limpieza del dataset sin afectar su representatividad.

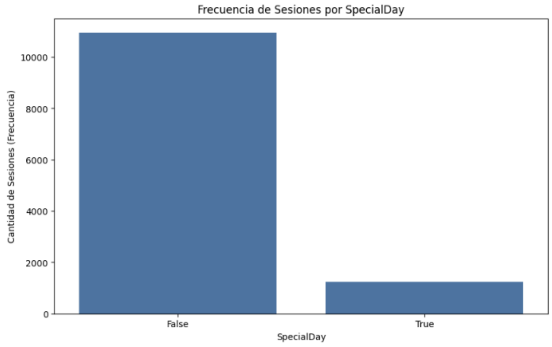

Figura 21*: mide la cercanía de la fecha de la visita a un día festivo. True si está cerca y False si no lo está.*

La variable **Month** representa el mes en el que se hizo la visita, al realizar una muestra de los valores correspondientes podemos observar que hay una inconsistencia con los otros valores almacenados, tal que el mes “June” no tiene 3 letras como el resto de los meses. Para el análisis, suponemos que si "June"

y "Jun" existieran serían dos categorías diferentes, lo cual es incorrecto. Por lo tanto, nuestra solución fue cambiar el valor “June” a “Jun” para que quede consistente con el resto de los valores del atributo Month.

```
Month
May      3364
Nov      2998
Mar      1907
Dec      1727
Oct       549
Sep       448
Aug       433
Jul       432
June      288
Feb       184
Name: count, dtype: int64
```

```
Month
May    3364
Nov    2998
Mar    1907
Dec    1727
Oct     549
Sep     448
Aug     433
Jul     432
Jun     288
Feb     184
Name: count, dtype: int64
```

Figura 22 y 23: *Se puede observar cómo queda en dataset antes y después del cambio.*

Además, se puede observar que el mes de mayo y noviembre fue donde hubo la mayor cantidad de visitas. Mientras que en el mes de febrero fue donde menos hubo la visitas. Pero si comparamos estos resultados con la variable “Revenue” podemos observar que el mes Noviembre (Nov) es el mes con el rendimiento más alto, logrando que un 25.4% de sus sesiones terminen en compra. Le siguen Octubre (Oct) con 21.0% y Septiembre (Sep) con 19.2%. En contraste, Mayo (May), a pesar de tener el mayor volumen de visitas, tiene un porcentaje bajo (10.9%) y Febrero (Feb) muestra un rendimiento casi nulo (1.6%).

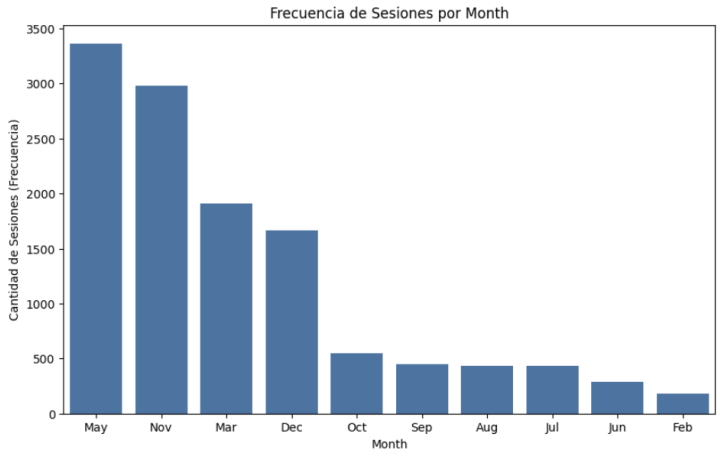

Figura 24: *Grafico de barras para representar los valores almacenados en la variable Month.*

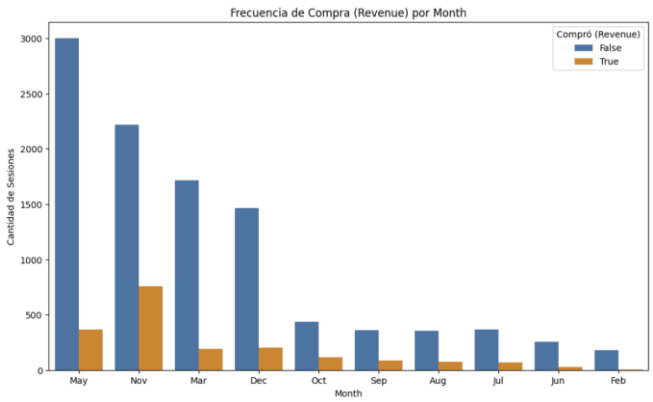

Figura 25: *Este grafico de barras compara las variables Month y Revenue, para observar la relación de compra con la visita del mes.*

La variable **VisitorType** define que tipo de visitante (reciente o nuevo) al realizar una muestra de los valores correspondientes podemos observar que hay una inconsistencia con los valores que podía almacenar, tal que en la descripción del atributo solo permitía que se guarde “Returning_Visitor” o “New_Visitor” y no un valor del tipo “Other”. Tal que consideramos que como esas 85 filas son solo el **0.69%** del dataset. Lo cual, es una fracción tan pequeña que su eliminación no moverá tus estadísticas generales (medias, medianas, etc.) de ninguna manera notable. Tal que, al eliminarlas, no estaríamos perdiendo información valiosa; estaríamos sacando ambigüedad, quedando así un dataset mucho más limpio, consistente y más fácil de interpretar.

```
VisitorType
Returning_Visitor    10551
New_Visitor           1694
Other                   85
Name: count, dtype: int64
```

```
VisitorType
Returning_Visitor    10551
New_Visitor           1694
Name: count, dtype: int64
```

Figura 26 y 27: *Representan los valores que tenía la variable antes y después de la eliminación.*

Además, se puede observar que, tras la limpieza, la variable VisitorType presenta un fuerte desbalanceo: la categoría "Returning_Visitor" representa el 85.5% de las sesiones, frente al 14.5% de "New_Visitor".

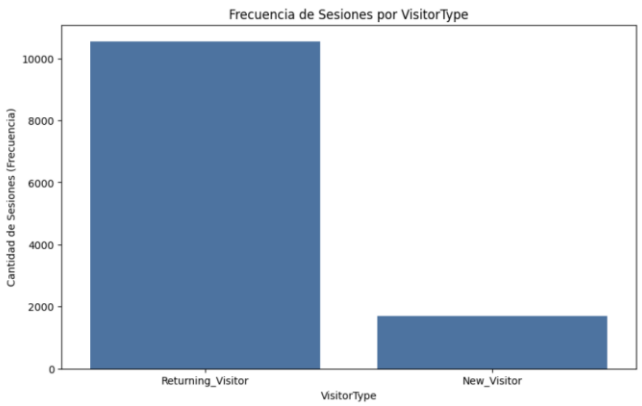

***Figura 28**: Representación estadística de la variable VisitorType.*

Al comparar esta variable con la intención de compra (Revenue), notamos que los visitantes nuevos ("New_Visitor") tienen una tasa de conversión (25.5%) significativamente más alta que los visitantes recurrentes ("Returning_Visitor", con 12.6%). Aunque la mayoría del tráfico es recurrente, los nuevos visitantes son comparativamente más propensos a comprar.

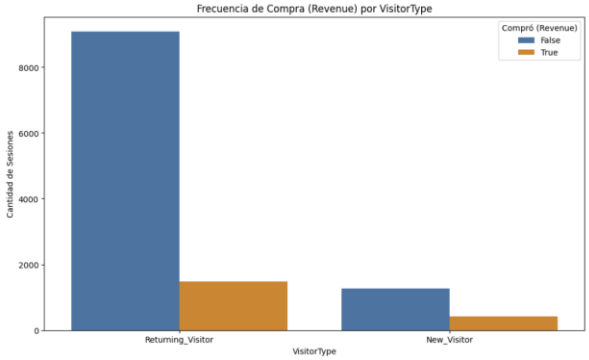

***Figura 29:** Representación estadística de la variable VisitorType y Revenue, tal que se compara que tipo de visitante termino en compra.*

La variable **OperatingSystem** representa que sistema operativo tiene el usuario en la visita, el cual está codificado en 8 categorías numéricas. El análisis de su distribución muestra una fuerte concentración en solo tres categorías: los sistemas 2 (53.5%), 1 (21.0%) y 3 (20.7%), que juntos suman más del 95% de todo el tráfico. Las categorías restantes (4, 5, 6, 7 y 8) son minoritarias y acumulan menos del 5% de las visitas.

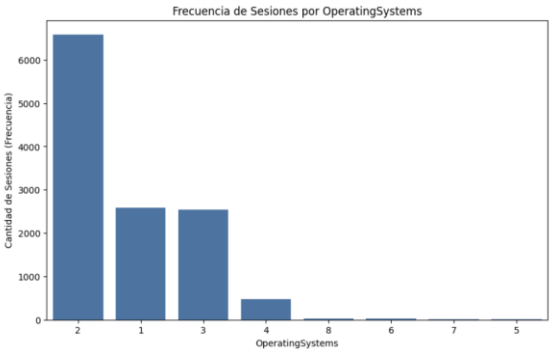

Figura 30: *El grafico representa el S.O que tiene el usuario en la visita*

Al analizar el gráfico que contiene la comparación entre la variable OperatingSystem y Revenue podemos ver que el tráfico está fuertemente concentrado en tres categorías: los sistemas 2 (53.5%), 1 (21.0%) y 3 (20.7%), sumando más del 95% de las visitas. Al observar los porcentajes podemos notar que no existe un patrón claro o una diferencia drástica que sugiera que un S.O sea significativamente mejor que otro. Los porcentajes de los tres grupos principales son 17.5% (SO 2), 14.7% (SO 1) y 10.5% (SO 3). Dado que estos son relativamente similares y no hay una tendencia clara, esta variable parece tener un bajo poder predictivo por sí sola.

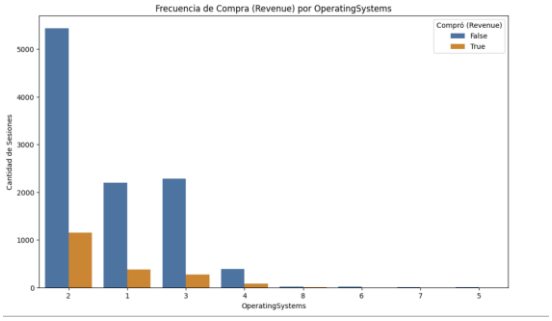

Figura 31: *El grafico compara la variable OperatingSystem y Revenue, para observar la relación que tiene el S.O utilizado por el usuario con la compra que haga el mismo.*

La variable **Browser** representa el navegador en el que se encuentra el usuario cuando se realiza la visita (teniendo 13 categorías numéricas) y la variable **TrafficType** (teniendo 20 categorías numéricas) representa cómo llegó el usuario al sitio.

Analizando el heatmap los navegadores 2 y 1 concentran la gran mayoría del tráfico. La altura total de la barra del Navegador 2 es significativamente mayor, indicando que es el más utilizado.

La composición interna de las barras (los colores) representa los diferentes TrafficType. A simple vista, se observa que los segmentos de color más grandes corresponden a las fuentes de tráfico más comunes: TrafficType 2 (3902 visitas), TrafficType 1 (2444) y TrafficType 3 (2045).

Lo más relevante es que la distribución de estos colores es muy similar en todas las barras principales. Tanto el Browser 1 como el Browser 2 obtienen sus visitas predominantemente de las mismas fuentes (1, 2 y 3). Esto sugiere que no existe una interacción o dependencia especial entre el navegador utilizado y la fuente de tráfico por la que llega un usuario.

La distribución de las fuentes de tráfico es muy similar en todos los navegadores principales. Tanto el Browser 1 como el Browser 2 (y los demás con menos tráfico) obtienen sus visitas predominantemente de los TrafficType 1, 2 y 3. Por lo tanto, no existe una interacción o dependencia especial entre el navegador utilizado y la fuente de tráfico por la que llega un usuario.

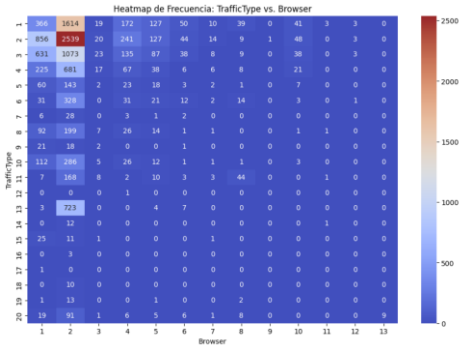

Figura 32: *El diagrama heatmap compara la variable Browser y TrafficType, tal que podemos observar la relación que hay entre ellas.*

La variable **Region** representa la zona desde la cual el usuario realiza la visita, teniendo 9 categorías numéricas.

El diagrama de barras (figura 32) muestra la distribución del tráfico y deja en evidencia un fuerte desbalanceo. La Región 1 es, con diferencia, la que más visitas genera (4772). Le siguen en importancia la Región 3 (2395) y la Región 4 (1177). En conjunto, estas tres regiones concentran la gran mayoría de las sesiones, mientras que las seis restantes (2,5,6,7,8 y 9) tienen un volumen de tráfico significativamente menor.

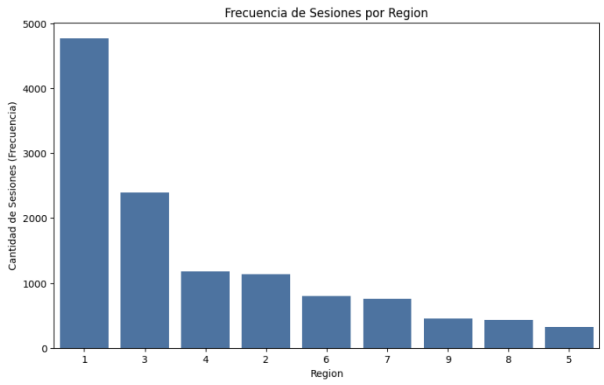

***Figura 33**: El diagrama de barras muestra la zona desde la cual el usuario realiza la visita.*

Analizando el diagrama (figura 33) podemos observar que el TrafficType 1 es la fuente más importante en general, generando tanto el mayor volumen de sesiones, como el mayor número absoluto de compras.

Sin embargo, el análisis muestra que la probabilidad de compra no es igual en todas las regiones. Aunque la Región 1 tiene la mayor cantidad de sesiones, su tasa de conversión es menor que la de regiones como la 2, 3 y 4, donde la proporción de compras respecto del total de visitas es más alta. En contraste, regiones de baja actividad (5, 8 y 9) presentan conversiones muy bajas.

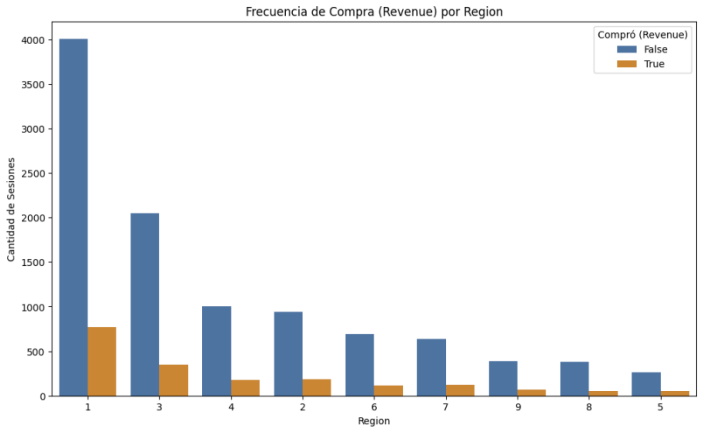

***Figura 34:** Grafico que muestra la relación entre la Region desde donde hace la visita el usuario y la variable Revenue para saber si este mismo realizo una compra (variable Revenue).*

La variable **Wekeend** indica si la visita fue en un fin de semana o no. Analizando el grafico es evidente que la gran mayoría del tráfico y de las sesiones de compra (en números absolutos) ocurren en días de semana (False). La altura total de la barra False es significativamente mayor que la de la barra True.

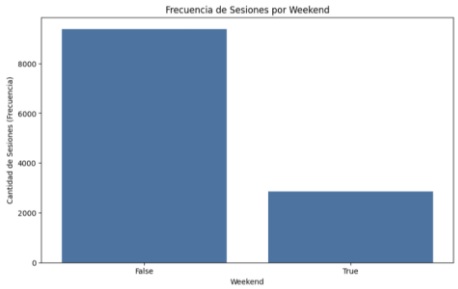

Figura 35: *El diagrama de barras muestra si la visita fue un fin de semana o no.*

El análisis muestra que, aunque durante los días de semana se concentra la mayor cantidad de sesiones, la tasa de conversión es proporcionalmente mayor durante el fin de semana. En términos relativos, el porcentaje de sesiones que terminan en compra es superior cuando es fin de semana en comparación con los días laborables. Esto sugiere que los usuarios que navegan durante el fin de semana presentan una mayor intención de compra.

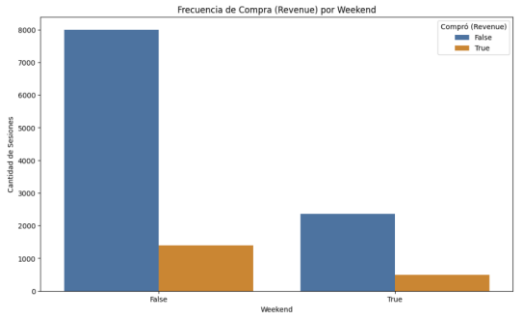

Figura 36: *El diagrama representa la relación entre la compra y la visita un fin de semana o no.*

La variable **Revenue** indica si la visita del usuario tuvo ganancias, es decir, si se realizó una compra. El gráfico evidencia una fuerte asimetría en la distribución: la gran mayoría de las sesiones no generan ingresos, mientras que solo una fracción relativamente pequeña termina en compra. Esta diferencia refleja una tasa de conversión baja, típica en entornos de comercio digitales.

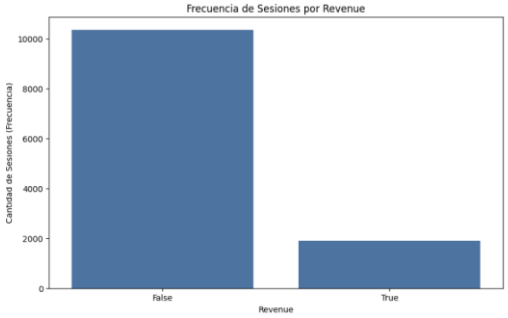

Figura 37: *El grafico representa si la vista del usuario termino en compra o no.*

2.2 Procesamiento de Outliers

Durante el análisis exploratorio se identificó la presencia de outliers en varias variables relacionadas con la duración de navegación y el número de páginas visitadas. Estos valores extremos se evidenciaron al comparar los percentiles superiores (especialmente el 75%) con los valores máximos, donde se observaron desviaciones excesivamente grandes que no representan un comportamiento realista de los usuarios.

Por ejemplo, en *ProductRelated_Duration* el percentil 75 es de aproximadamente 1477 segundos, mientras que el valor máximo registrado supera las 17 horas. De manera similar, *Informational_Duration* muestra un percentil 75 igual a 0 pero un máximo superior a 40 minutos. Estas diferencias sugieren errores de registro o sesiones anómalas, por lo que su presencia puede distorsionar el análisis estadístico y cualquier modelo posterior.

	count	mean	std	min	25%	50%	75%	max
Administrative	12119.0	2.345160	3.334828	0.0	0.000000	1.000000	4.000000	27.000000
Administrative_Duration	12119.0	81.785984	177.300976	0.0	0.000000	9.500000	95.000000	3398.750000
Informational	12119.0	0.511098	1.278520	0.0	0.000000	0.000000	0.000000	24.000000
Informational_Duration	12119.0	34.990626	141.824727	0.0	0.000000	0.000000	0.000000	2549.375000
ProductRelated	12119.0	32.185411	44.667644	0.0	8.000000	18.000000	38.000000	705.000000
ProductRelated_Duration	12119.0	1211.546867	1922.747900	0.0	196.416667	614.640952	1483.411905	63973.522230
BounceRates	12119.0	2.022798	4.488808	0.0	0.000000	0.298507	1.666667	20.000000
ExitRates	12119.0	4.129903	4.588686	0.0	1.426082	2.500000	4.819778	20.000000
PageValues	12119.0	5.864200	18.159499	0.0	0.000000	0.000000	0.000000	361.763742

Figura 38: *Serie de estadísticas descriptivas de todas las columnas, ya sea el desvió estándar (std), valor mínimo de esa columna (min), etc.*

Con el objetivo de preservar la mayor cantidad de información sin eliminar observaciones completas, se optó por aplicar un *método de capping*. Este procedimiento consiste en reemplazar únicamente los valores ubicados por encima del *percentil 99* por el valor correspondiente a dicho percentil. De esta forma, se suavizan las colas de las distribuciones sin afectar el comportamiento general de la variable ni introducir sesgos significativos.

Las variables a las que se les aplicó este tratamiento fueron:

- *Administrative*
- *Administrative_Duration*
- *Informational*
- *Informational_Duration*

- *ProductRelated*
- *ProductRelated_Duration*

Estas variables presentaban valores máximos claramente desproporcionados y alejados de su rango intercuartílico, lo que justificó su corrección.

En contraste, variables como **BounceRates**, **ExitRates** y **PageValues** no recibieron ningún tratamiento. En estos casos, los valores altos forman parte natural de su comportamiento estadístico: las tasas pueden tomar valores elevados sin ser errores, y *PageValue* es una métrica económica donde los montos elevados reflejan sesiones con alta contribución al proceso de compra. Por lo tanto, sus valores extremos contienen información válida y no deben ser modificados.

	count	mean	std	min	25%	50%	75%	max
Administrative	12119.0	2.315290	3.202912	0.0	0.000000	1.000000	4.000000	14.000000
Administrative_Duration	12119.0	77.185506	142.712723	0.0	0.000000	9.500000	95.000000	837.672000
Informational	12119.0	0.492697	1.156620	0.0	0.000000	0.000000	0.000000	6.000000
Informational_Duration	12119.0	30.836692	105.541565	0.0	0.000000	0.000000	0.000000	722.831600
ProductRelated	12119.0	31.261078	38.372786	0.0	8.000000	18.000000	38.000000	221.000000
ProductRelated_Duration	12119.0	1168.925364	1557.227130	0.0	196.416667	614.640952	1483.411905	8703.918663
BounceRates	12119.0	2.022798	4.488808	0.0	0.000000	0.298507	1.666667	20.000000
ExitRates	12119.0	4.129903	4.588686	0.0	1.426082	2.500000	4.819778	20.000000
PageValues	12119.0	5.864200	18.159499	0.0	0.000000	0.000000	0.000000	361.763742

***Figura 39**: Valores estadísticos descriptivas de las variables luego de aplicarles el método capping.*

Este tratamiento permite conservar la estructura general del dataset, reducir la influencia de puntos atípicos y mejorar la calidad de los análisis posteriores sin comprometer la integridad de la información.

3. Hipótesis planteadas y resolución

3.1. Hipótesis 1:

3.1.1. Definición de la hipótesis

El objetivo de esta hipótesis es analizar si el mes de mayo presenta una proporción de visitas distinta al 30% del total anual.
Este análisis es relevante porque mayo es, visualmente, el mes con mayor cantidad de visitas en el dataset, por lo que resulta importante validar estadísticamente si dicho aumento implica superar un porcentaje “alto” (30%).

- **H0 (hipótesis nula):** La proporción de visitas en mayo es **igual o menor** al 30%.
- **H1 (hipótesis alternativa):** La proporción de visitas en mayo es **mayor** al 30%.

Esta hipótesis busca confirmar si mayo puede considerarse un mes de “pico” respecto al resto del año o no alcanza un 30% del total anual, aunque sea el mes con más visitas.

3.1.2. Estrategia de abordaje

Ya que la hipótesis planteada busca comparar la proporción de visitas que ocurren en un mes (mayo) contra un valor teórico (30%). se realizó una prueba de proporciones (test Z), comparando:

- La proporción observada en mayo respecto al total anual,
- Valor poblacional hipotético de 0.30.

El procedimiento fue:

1. **Identificar valores:**
    - Total, de visitas en mayo: **3325**
    - Total, de visitas anuales: **12119**
    - Proporción observada: **p=0.2744 (27.44%)**
2. **Plantear la prueba unilateral (cola derecha)**:
   Se evalúa si la proporción real supera el 30%.
3. **Cálculo del estadístico Z:**

Z= -6.1588

Obtención del p-value:

p-value=1−Φ(Z)

P-value≈0.9999999998 Este valor extremadamente cercano a 1 indica que la proporción observada está muy por debajo del 30%, en contradicción directa con la hipótesis alternativa.

3.1.3. Resultados obtenidos y discusión

El resultado estadístico indica que:

- El estadístico Z =-6.1588 es muy negativo, lo que significa que la proporción observada (27.44%) está muy por debajo de la proporción hipotética de 30%.
- El p-value unilateral ≈ 1 implica que no existe evidencia estadística para rechazar H0.

No se confirma la hipótesis de que la proporción de visitas en mayo sea mayor al 30%. En lugar de ello, los datos muestran claramente que mayo, si bien es el mes con más visitas absolutas del año, no alcanza el umbral del 30%, quedando aproximadamente 3 puntos porcentuales por debajo.La hipótesis H1 se rechaza, la proporción de visitas en mayo no es mayor al 30% y es significativamente menor.

Apoyo visual:

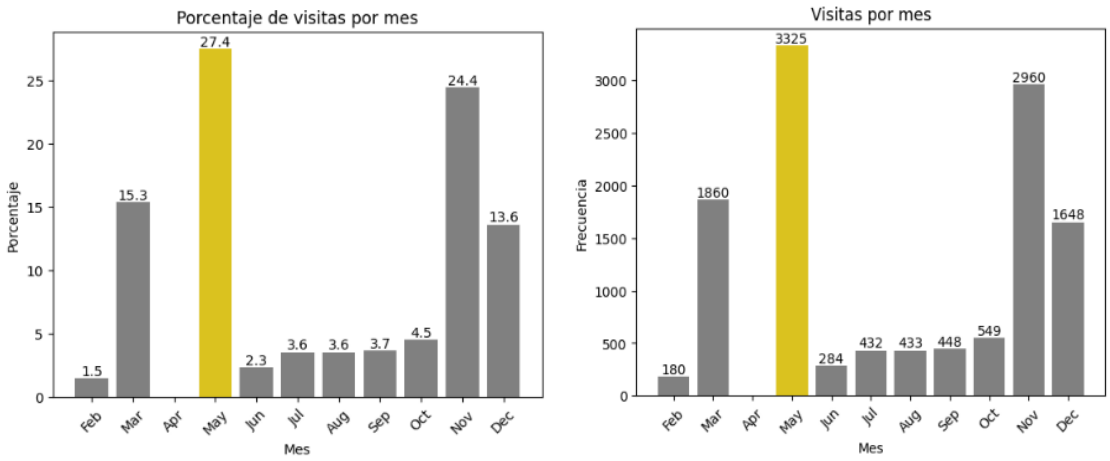

Figura 40 y 41: *representa las cantidades de visitas por mes junto a su porcentaje*

3.2. Hipótesis 2:

3.2.1. Definición de la hipótesis

El objetivo de esta hipótesis es analizar si existe una relación entre dos variables categóricas del dataset (Tipo de cliente, Ganancia)

Esto permite determinar si ciertos tipos de clientes generan más o menos ganancia que otros, lo cual es importante para decisiones comerciales.

Las hipótesis planteadas son:

- **H0 (Hipótesis nula):** El tipo de cliente es **independiente** de la ganancia (No existe relación entre ambas variables).
- **H1 (Hipótesis alternativa):** El tipo de cliente **no es independiente** de la ganancia (Sí existe relación entre tipo de cliente y ganancias).

3.2.2. Estrategia de abordaje

Como ambas variables son categóricas, la técnica que usamos es la prueba Chi-cuadrado de independencia, esta prueba evalúa si la distribución observada de frecuencias entre dos variables categóricas difiere significativamente de la que se esperaría bajo independencia.

Construcción de la tabla de contingencia (observados)

Observado:

- (Fila 1, 1): **1271**
- (Fila 1, 2): **422**
- (Fila 2, 1): **8956**
- (Fila 2, 2): **1470**

Cálculo de frecuencias esperadas bajo independencia: (total fila)(total columna) /N

Esperada:

- (Fila 1, 1): **1428.69**
- (Fila 1, 2): **264.30**
- (Fila 2, 1): **8798.30**
- (Fila 2, 2): **1627.70**

Cálculo del estadístico Chi^2:

Sumando: $\chi 2 = 128.769 \approx 128$

Grados de libertad:

$gl=(r-1)(c-1)=(2-1)(2-1)= 1$

Valor crítico

$\alpha = 0.05$:

$\chi_{crítico2} = 3.841$ (Valor critico de tabla para gl=1)

Comparación: Estadístico: $128 \gg 3.841$

p-value: Extremadamente pequeño:$p<0.00001$

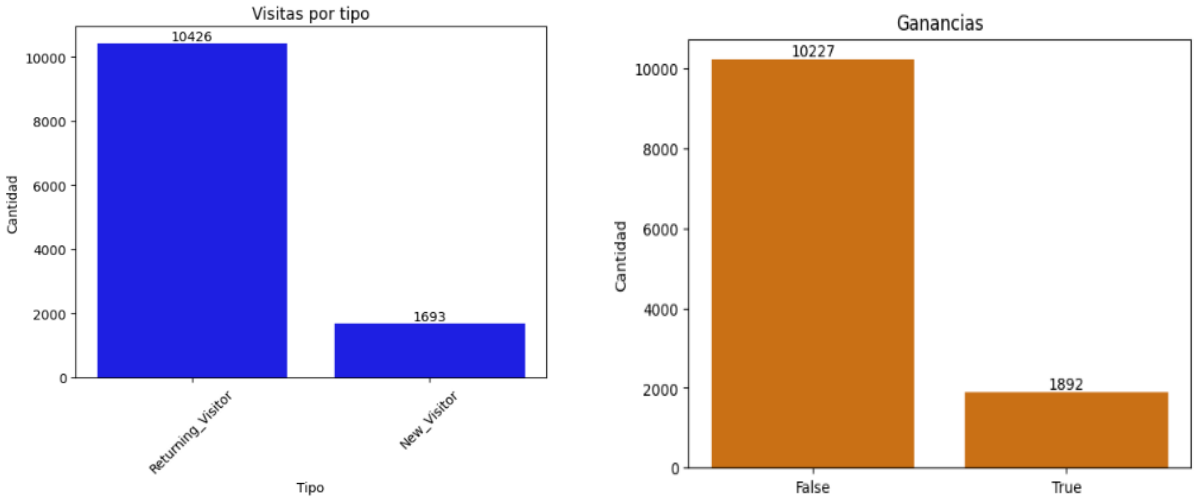

Figura 42 y 43: *Cantidad de visitas por tipo y la cantidad de estas que tuvieron ganancias*

3.2.3. Resultados obtenidos y discusión

El valor Chi^2 calculado (≈ 128) es mucho mayor que el valor crítico (3.841).

También el p-value < 0.00001, lo que indica que la probabilidad de obtener diferencias tan grandes entre las frecuencias observadas y esperadas es prácticamente nula.

Se rechaza H_0 las variables no son independientes porque existe una relación estadísticamente significativa entre el tipo de cliente y la ganancia generada.

Se podría decir que:

- Algunos tipos de clientes aportan ganancias mayores o menores.
- El comportamiento de ganancia depende del tipo de cliente.
- Los clientes no se distribuyen al azar entre niveles de ganancia.

Ayuda visual:

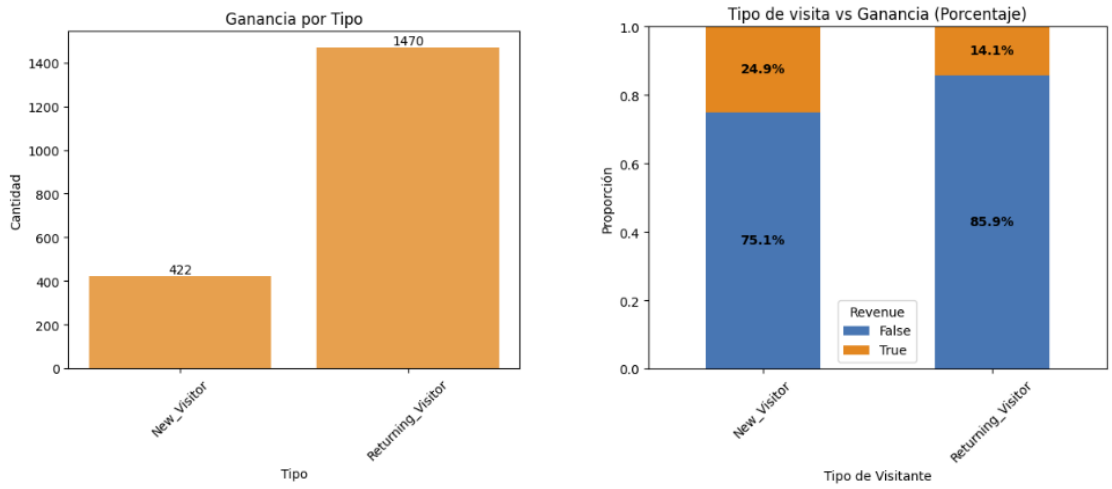

Figura 44 y 45: *Cantidad de ganancias por tipo de visitante y el porcentaje de ganancia o no dentro de cada tipo*

3.3. Hipótesis 3:

3.3.1. Definición de la hipótesis

Durante el análisis exploratorio observamos que la tasa de rebote (Bounce Rate) presentaba una media cercana al 2%, pero con cierta variabilidad entre visitas. Dado que el Bounce Rate es un indicador clave del comportamiento del usuario resulta relevante determinar si su promedio supera o no el umbral estándar del 2%.

H0 (hipótesis nula): El promedio del Bounce Rate es menor o igual al 2%.

H1 (hipótesis alternativa): El promedio del Bounce Rate es mayor al 2%.

En la industria del marketing digital, un Bounce Rate del 2% suele considerarse un valor de referencia bajo, por lo que un promedio que realmente lo supere podría indicar problemas en la relevancia del contenido o en la navegación del sitio.

3.3.2. Estrategia de abordaje

Para contrastar estas hipótesis se empleó un test t de una muestra con cola derecha, ya que:

1. La variable BR es continua y se estudia su media.
2. Se compara contra un valor fijo de referencia (0.02).

A partir de los datos del Bounce Rate se obtuvo:

- Tamaño muestral: 12119
- Media muestral: 2.0228
- Desviación estándar: 4.4888
- Estadístico t: 0.5591
- p-value unilateral: 0.288050

3.3.3. Resultados obtenidos y discusión

El p-value se obtuvo utilizando la función de distribución de la t-Student con 1 grados de libertad. Dado que el p-value = 0.288050 es mayor que α = 0.05, no se rechaza la hipótesis nula.

No hay evidencia suficiente para afirmar que el promedio del Bounce Rate supera el umbral del 2% entonces:

- No se rechaza H_0 según los datos disponibles. La tasa de rebote del sitio es estadísticamente inferior al 2%, lo que sugiere que el comportamiento general de los usuarios se encuentra dentro del rango esperado.

Ayuda visual:

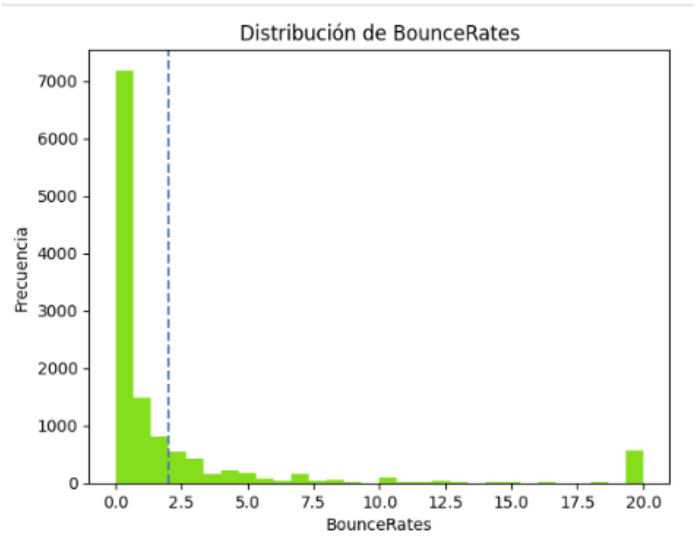

Figura 46: *Histograma con la distribución de BR*

3.4. Hipótesis 4:

3.4.1. Definición de la hipótesis

Planteamos la siguiente hipótesis bivariada entre las variables SpecialDay (booleana) y ProductRelated (numérica):

Dado que intuitivamente los días especiales podrían atraer más visitantes, decidimos formular la hipótesis:

- **H0 (Hipótesis nula):** Los días especiales NO tienen mayor cantidad de visitas en productos que los días normales.
- **H1 (Hipótesis alternativa):** Los días especiales tienen mayor cantidad de visitas en productos que los días normales**.**

El potencial de esta hipótesis radica en evaluar si los días especiales influyen en el interés de los usuarios por explorar productos, lo cual podría ser relevante para decisiones comerciales.

3.4.2. Estrategia de abordaje

Dado que:

- SpecialDay divide la muestra en dos grupos independientes (0 y 1)
- ProductRelated es una variable numérica continua

la prueba estadística adecuada es una prueba t para muestras independientes (compara las medias de dos grupos no relacionados y determina si hay una diferencia estadísticamente significativa entre ellos).

El resultado se compara con una distribución t de Student para obtener el p-valor, que indica si la diferencia observada es estadísticamente significativa bajo H0 .

El nivel de significancia utilizado fue $\alpha = 0.05$.

3.4.3. Resultados obtenidos y discusión

El test produjo los siguientes resultados:

- t = -4.5633
- p-valor = 5.39 × 10^-6

El p-valor es mucho menor que 0.05, por lo que corresponde rechazar la hipótesis nula (H_0).

El valor t es negativo, lo cual indica que los usuarios en días especiales visitaron menos páginas de productos que en días normales.

Los resultados del análisis permiten rechazar la hipótesis nula (H_0), dado que el p-valor obtenido es muy inferior al nivel de significancia establecido. Esto indica que sí existe una relación entre la condición de día especial (SpecialDay) y la cantidad de páginas de productos visitadas (ProductRelated) pero, aunque la relación existe, la dirección del efecto es opuesta por lo que la hipótesis alternativa (H_1) se confirma:

Los días especiales no generan un aumento significativo en la cantidad de visitas en productos.

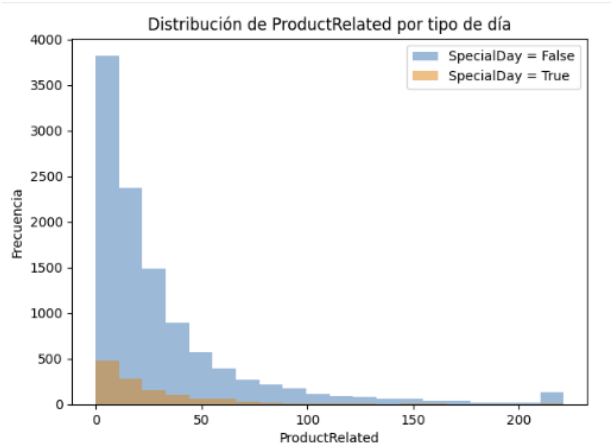

Figura 47: *Histograma con la proporción de visitas (ProductRelated) si fueron días especiales y las que no*

3.5 Hipótesis 5:

3.5.1. Definición de la hipótesis

Esta hipótesis se basa en la idea de que podemos segmentar a los usuarios en "interesados" (visitantes que invierten tiempo en el sitio y con baja tasa de rebote) vs. "perdidos" (visitantes que invierten poco tiempo en el sitio y con alta tasa de rebote) usando estas dos métricas de comportamiento opuestas planteamos la siguiente hipótesis multivariada entre las variables siguientes variables:

X: (Variables Independientes): Columnas ['ProductRelated_Duration', 'BounceRates'].

Y: (Variable Dependiente): ['Revenue'].

- **H0 (hipótesis nula):** No existe una relación significativa entre el crecimiento de ProductRelated_Duration y el decrecimiento de BounceRates con Revenue.
- **H1 (hipótesis alternativa):** "Existe una relación en cuanto al crecimiento de ProductRelated_Duration (tiempo en productos) combinado con un bajo BounceRates (tasa de rebote) que predecirá una probabilidad mucho mayor de Revenue (compra)."

3.5.2. Estrategia de abordaje

Analizamos el comportamiento de las variables ProductRelated_Duration y BounceRates contra la variable Revenue mediante gráficos y su promedio.

Análisis mediante gráficos:

Los análisis fueron uno a uno (ambas variables por separado con Revenue) y uno en conjunto (distribución de Revenue)

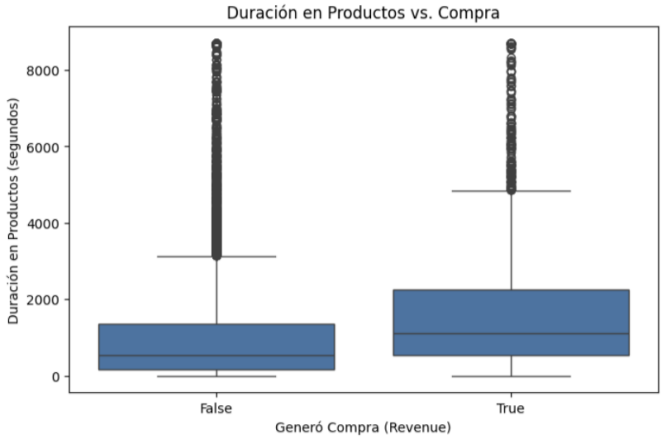

Figura 48: *Box-plot ProductRelated_Duration con respecto a los valores de Revenue*

Caja False: La línea del medio (mediana) está muy abajo.

Caja True: La línea del medio está visiblemente más alta que la línea del False.

La gente que compró (True) tiene una mediana de tiempo mucho más alta (pasó más tiempo en productos) que la gente que no compró (False).

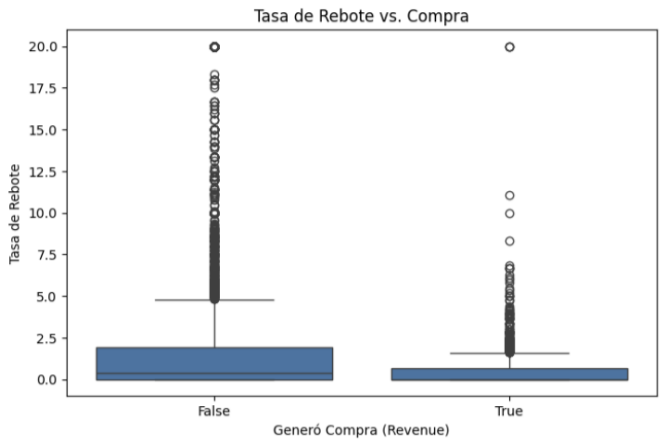

Figura 49: *Box-plot BounceRates (%) con respecto a los valores de Revenue*

Caja False: La caja es "gorda", va de 0 a casi 2.5%, y su línea del medio (mediana) está por encima de 0.

Caja True: La caja está completamente aplastada contra el 0, es una más fina.

La gente que compró (True) tiene una mediana de tasa de rebote de casi 0. En cambio, la gente que no compró (False) tiene tasas de rebote mucho más altas y variables.

Ambos gráficos muestran Indicios de que la hipótesis multivariada va por buen camino, observemos cómo se comportan las variables juntas:

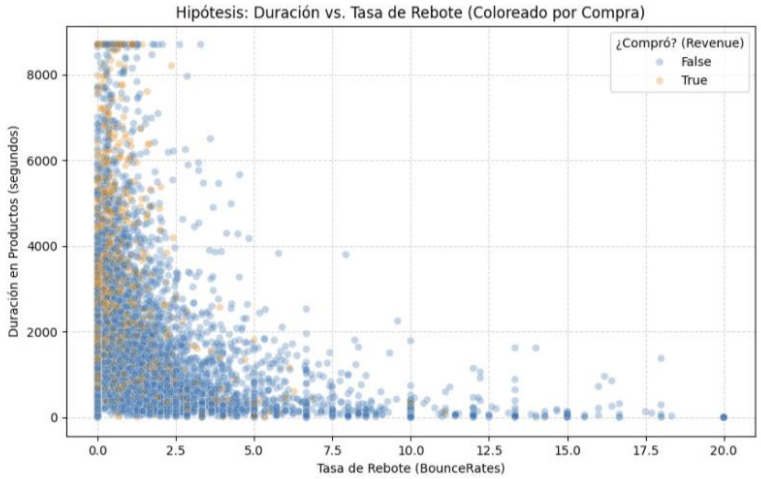

Figura 50: *Scatter-Plot distribución de revenue's con respecto a ProductRelated_Duration y BounceRates (%)*

Eje X (Izquierda a Derecha): Tasa de Rebote (BounceRates).

Eje Y (Abajo a Arriba): Duración en Productos (segundos).

El Color (El Resultado): ¿Compró? (Revenue)

- Puntos Celestes (False): Sesiones que NO terminaron en compra.
- Puntos Naranjas (True): Sesiones que SÍ terminaron en compra.

Análisis figura 50:

- **En la izquierda** (Tasa de Rebote de 0.0 a 2.5%). Está lleno de puntos celestes (False) Y naranjas (True). Están todos mezclados. Pero si prestamos atención a los puntos naranjas son los que de manera agrupada se encuentran más arriba (mayor duración de tiempo en página de productos y baja tasa de rebote). Sin embargo, los puntos azules que están en la misma zona del eje x tienden a estar en grupo más hacia abajo (es decir cuentan con poco tiempo en la página).
- **En la derecha** (Tasa de Rebote > 5% o incluso > 2.5%). No se ven prácticamente puntos naranjas cuando la tasa de rebote aumenta.

Por lo tanto, se puede observar que, si la tasa de rebote es alta, la probabilidad de compra desaparece y que a menor tiempo de duración se compra menos.

Análisis mediante promedio:

Separamos a las +12000 sesiones en dos grupos (los que compraron True y los que no False) para luego calcular el promedio de las 2 variables para cada grupo.

Revenue	ProductRelated_Duration	BounceRates
False	1051.734154	2.301693
True	1802.389686	0.515260

Figura 51: *Agrupación por Revenue, promedio de ProductRelated_Duration y BounceRates.*

ProductRelated_Duration: El grupo que compró (True) pasó, en promedio, casi el doble de tiempo mirando productos. Buen indicativo para la parte de la hipótesis, Alta Duración = Compra.
BounceRates: El grupo que compró (True) tuvo, en promedio, una tasa de rebote ~5 veces menor. Esto es otro buen indicativo para la otra parte de la hipótesis, Bajo Rebote = Compra.

3.5.3. Resultados obtenidos y discusión

Validación:

Como es una hipótesis multivariada que observa si la variable dependiente (Revenue) es afectada por las 2 variables independientes ('ProductRelated_Duration', 'BounceRates'), utilizaremos una validación por regresión.

	Coeficiente (Peso Estadístico)
ProductRelated_Duration	0.352155
BounceRates	-1.537734

Precisión de la Validación: 0.841

Figura 52: Coeficientes estadísticos.

ProductRelated_Duration: 0.352155 (Positivo +):

El coeficiente es positivo. Esto significa que, manteniendo todo lo demás constante, un aumento en el tiempo que alguien pasa en páginas de producto aumenta la probabilidad de que compre.

BounceRates: -1.537734 (Negativo -):

El coeficiente es negativo. Esto significa que un aumento en la tasa de rebote disminuye drásticamente la probabilidad de que alguien compre.

Además, se puede observar en la que magnitud del coeficiente en BounceRates (-1.537734) en contra del coeficiente en ProductRelated_Duration (+0.352155) es mayor, esto quiere decir que la inferencia de BounceRates en Revenue es mucho mayor.

Los resultados del modelo de Regresión Logística validan estadísticamente la hipótesis H1, ya que se encontró una correlación positiva (+0.352155) para ProductRelated_Duration y una correlación negativa fuerte (-1.537734) para BounceRates sobre la probabilidad de Revenue.

La fiabilidad de este hallazgo se ve respaldada por una precisión de validación del 84.1%.

3.6. Hipótesis 6:

3.6.1. Definición de la hipótesis

En esta hipótesis analizamos si PageValues es distinto entre quienes terminan comprando y quienes no. PageValues es una métrica de Google Analytics que indica cuánto “valor económico” aportan las páginas que visitó un usuario. Por cómo funciona el sitio, es esperable que las sesiones que llegan a zonas clave (productos, carrito, checkout) tengan comportamientos distintos a las que no.

Planteamos:

- **H0 (hipótesis nula):** no hay diferencias en la distribución (mediana) de PageValues entre compradores y no compradores.
- **H1 (hipótesis alternativa):** La distribución de PageValues **es diferente** entre las sesiones que realizan una compra y las que no.

3.6.2. Estrategia de abordaje

Como se buscaba comparar la distribución de PageValues entre dos grupos independientes (compran vs. no compran), primero se verificaron los supuestos de las pruebas paramétricas. El test de *Shapiro Wilk* aplicado a una muestra representativa de cada grupo indicó que ambos presentan distribuciones altamente no normales ($p < 0.001$). Además, el test de *Levene* mostró diferencias significativas entre las varianzas ($p < 0.001$).

Debido a que no se cumplen los supuestos de normalidad ni de homogeneidad de varianzas, se optó por utilizar una prueba no paramétrica para dos muestras independientes. En particular, se aplicó el test *Mann Whitney U* en su forma estándar (bilateral), que permite comparar las distribuciones sin asumir normalidad.

Ayuda visual:

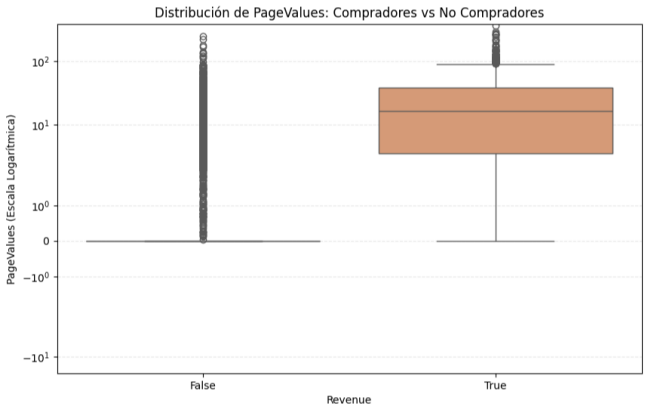

***Figura 53**: Boxplot que muestra la distribuciónde PageValues para sesiones con y sin compra. Se observa que las sesiones que concretan una transacción presentan valores significativamente más altos.*

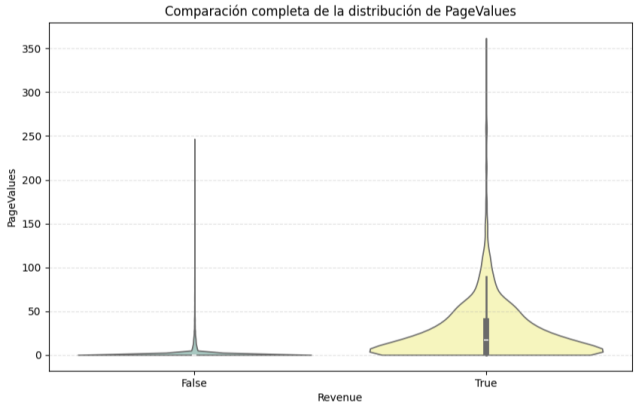

***Figura 54:** Gráfico de violín que representa la forma completa de la distribución de PageValues según Revenue. Las sesiones con compra muestran una concentración mayor de valores elevados.*

3.6.3. Resultados obtenidos y discusión

Los estadísticos descriptivos mostraron una clara diferencia inicial: las sesiones sin compra presentan una mediana de PageValues igual a 0, mientras que las sesiones con compra tienen una mediana de 16.67, lo que sugiere una tendencia en la dirección esperada.

Para validar la elección del test, se evaluaron los supuestos:

- **Normalidad:** El test de Shapiro-Wilk arrojó p-valores extremadamente bajos para ambos grupos ($p<0.001$), confirmando que los datos no siguen una distribución normal.
- **Homocedasticidad:** El test de Levene indicó varianzas no homogéneas ($W=3200.4$, $p<0.001$).

Dado el incumplimiento de estos supuestos, se procedió con el Test U de Mann-Whitney (bilateral). El resultado arrojó un estadístico $U=16,682,680$ y un p-valor < 0.001.

Dado que el p-valor es inferior al nivel de significancia ($\alpha=0.05$), rechazamos la Hipótesis Nula. Existe evidencia estadística suficiente para afirmar que las distribuciones de PageValues difieren significativamente entre quienes compran y quienes no. Combinando esto con el análisis descriptivo (mediana 16.67 vs 0), concluimos que los usuarios que realizan una compra navegan por páginas de mayor valor económico. Este resultado es coherente con el funcionamiento del sitio: las sesiones que avanzan más en el proceso de compra visitan páginas de mayor valor económico, mientras que la mayoría de las sesiones sin compra permanecen en valores cercanos a cero.

4. Conclusiones

El trabajo permitió comprender en profundidad los factores que influyen en la compra dentro del dataset “Online Shoppers Purchasing Intention”. A partir del análisis, el preprocesamiento y la validación de las hipótesis estadísticas, se consiguieron resultados que ayudan a evaluar el comportamiento de los usuarios en un entorno de e-commerce.

En primer lugar, el análisis exploratorio permitió detectar problemas estructurales (tipos de datos incorrectos, duplicados, outliers y categorías inconsistentes) que hubieran afectado directamente la calidad de los resultados. Estas transformaciones fueron importantes para obtener un dataset más representativo del comportamiento real de los usuarios.

En cuanto a las hipótesis, los resultados muestran patrones de la lógica del proceso de compra:

- H1 demostró que, aunque mayo es el mes con mayor cantidad de visitas absolutas, no alcanza el 30% del total anual. Por lo tanto, no puede considerarse un "pico" en términos proporcionales.
- H2 confirmó que sí existe una relación significativa entre el tipo de visitante y la probabilidad de compra.
- H3 mostró que el Bounce Rate promedio no supera el umbral del 2%, indicando un comportamiento positivo en la navegación.
- H4 permitió determinar que los días especiales no aumentan la cantidad de visitas a páginas de productos.
- H5, enfoque multivariado, confirmó que un mayor tiempo en páginas de productos combinado con una baja tasa de rebote está fuertemente asociado con la compra. Ambos factores (con mayor predominancia en BounceRates) demostraron ser relevantes para comprender el comportamiento final del usuario.
- H6 estableció que las distribuciones de PageValues entre compradores y no compradores son significativamente distintas: las sesiones que terminan en compra visitan páginas de mayor valor económico, lo cual es consistente con que avanzan más lejos en el proceso de conversión.

En conclusión, el estudio permitió identificar qué variables tienen verdadero impacto sobre la compra y cuáles no, aportando evidencia estadística clara para la toma de decisiones. El análisis confirma que los comportamientos dentro del sitio son el principal indicador de intención de compra en el e-commerce.

Referencias

- Enlace del dataset:
  https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset
- Documentación de scipy: https://docs.scipy.org/doc/scipy
- Documentación de Numpy: https://numpy.org/doc/stable
- Documentación entregada por la catedra:
  https://moodle.exa.unicen.edu.ar/course/view.php?id=1342&section=1#tabs-tree-start