Trabajo Práctico Especial

Fundamentos de Ciencias de Datos

2025

Integrantes: Martina Goncalves Dias

Santiago Orona

Christian Scornaienchi

1. Introducción

El dataset utilizado contiene un análisis en tiempo real de las acciones de compradores online. El mismo consiste de dos módulos, los cuales predicen la acción de compra del visitante y el abandono del sitio web, y la intención de compra usando clickstream e información de sesión, de manera simultánea.

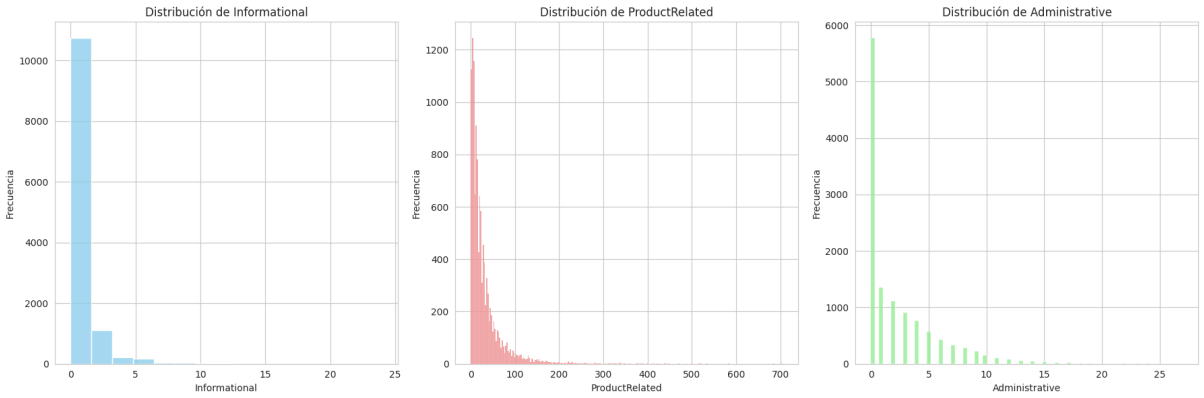
Se presentan varios datos, como "Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" y "Product Related Duration". Estos representan el número de los distintos tipos de páginas visitadas por un usuario en una sesión, y el total de tiempo que permaneció en cada una de estas, por categoría. Los valores de estas categorías son extraídos del URL de las páginas visitadas por el usuario y subidos en tiempo real cuando un usuario realiza una acción dentro de la página, por ejemplo moverse de una página a otra.

Durante el desarrollo del trabajo, se observaron distintos patrones. Por ejemplo, se detectó que en la variable Administrative Duration hay una mayor cantidad de valores iguales a cero en comparación con la variable Administrative. Esto puede interpretarse de distintas formas: el usuario pudo haber pasado un tiempo demasiado breve para ser registrado, o pudo haberse producido un error de conexión durante la carga de la página.

Este informe se organiza de la siguiente manera: en la Sección 2 se presenta el análisis exploratorio del conjunto de datos, describiendo los tipos de variables. En la Sección 3, se formularán las hipótesis derivadas de las observaciones previas y se detallarán las estrategias de análisis empleadas para su comprobación. Finalmente, en la Sección 4, se expondrán las conclusiones generales obtenidas a partir del estudio.

2. Análisis Exploratorio de los Datos

Las variables Administrative, Informational y Product Related representan los tres tipos de páginas que un usuario puede visitar dentro del sitio: administrativas, informativas y relacionadas con productos, respectivamente. Estas variables son cuantitativas discretas, ya que expresan la cantidad de páginas de cada tipo visitadas en una sesión.

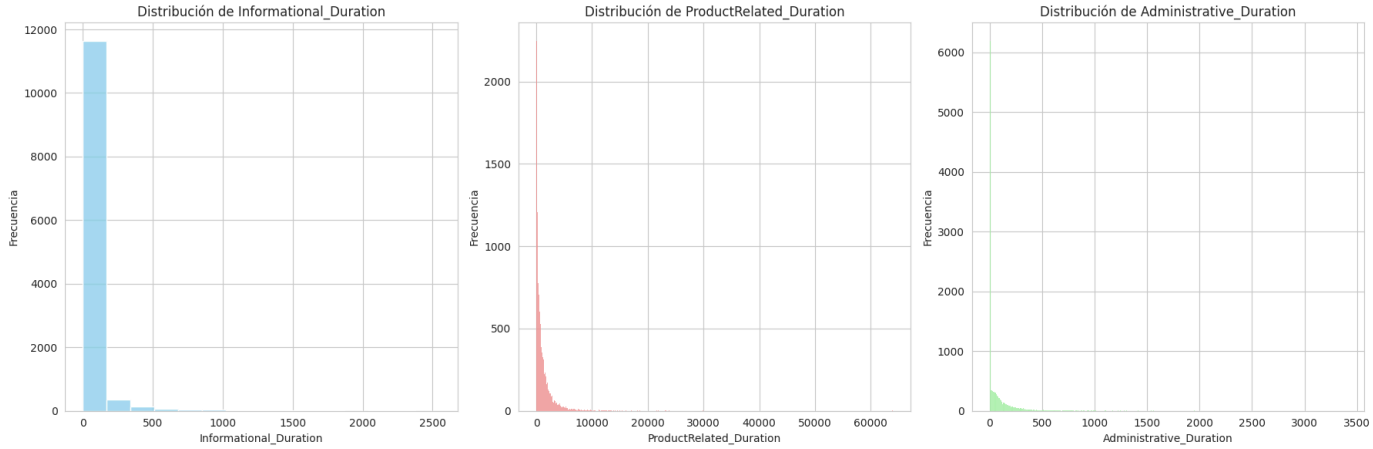


La distribución de las visitas a páginas informativas muestra un patrón fuertemente concentrado en valores bajos, con una gran proporción de sesiones en las que los usuarios no acceden a este tipo de contenido.Aunque existen algunos valores altos, estos corresponden a casos aislados y no representan una tendencia dominante. En general, este comportamiento sugiere que la información adicional del sitio no es un punto clave durante la navegación típica del usuario, ya sea porque no es necesaria para completar sus objetivos o porque no se encuentra suficientemente integrada en el recorrido del cliente.

Las páginas administrativas presentan también una distribución sesgada hacia la derecha, aunque con mayor diversidad respecto a las páginas informativas. Si bien hay un número considerable de sesiones sin visitas a este tipo de páginas, los valores más altos son algo más frecuentes. Esto indica que, aunque la mayoría de los usuarios no necesitan interactuar con funcionalidades administrativas durante su recorrido, existe un segmento menor que realiza acciones más complejas, como configuraciones, consultas de políticas o procesos adicionales dentro del sitio. Esta variabilidad sugiere que este tipo de páginas cumple un rol más específico y orientado a necesidades puntuales, propias de usuarios avanzados o con requerimientos particulares.

En contraste con las otras categorías, la distribución de visitas a páginas relacionadas con productos revela una interacción mucho más intensa y heterogénea. Aunque la distribución sigue siendo asimétrica, presenta una mayor dispersión y valores significativamente más altos, con un rango que llega hasta más de 700 páginas visitadas en una sola sesión. La mediana de 18 y una media de 31 reflejan que la mayoría de los usuarios exploran varios productos en cada visita, haciendo de esta categoría el eje central de la experiencia del sitio. La presencia de usuarios que navegan una cantidad excepcionalmente alta de páginas sugiere comportamientos de exploración intensiva, posiblemente vinculados a procesos de comparación o decisiones de compra más elaboradas. En conjunto, este patrón confirma la importancia de las páginas de productos como motor principal de interacción dentro del sitio.

En relación a las columnas mencionadas anteriores encontramos Administrative Duration, Informational Duration y Product Related Duration reflejan el tiempo que estuvo el usuario dentro de una pagina en cada tipo. Se tratan de variables cuantitativas continuas, medidas en segundos (por suposición).

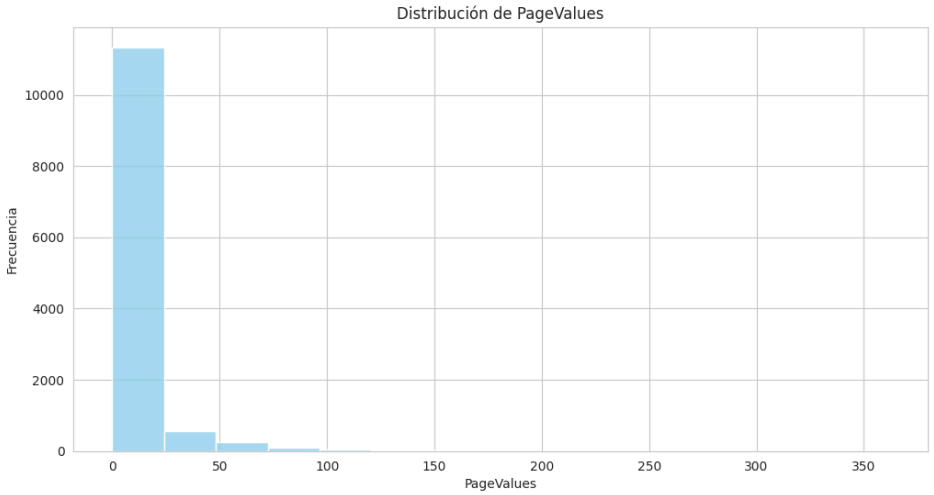


El tiempo de permanencia en páginas informativas muestra una distribución altamente concentrada cerca de cero, con una gran cantidad de sesiones en las que el usuario apenas interactúa con este tipo de contenido. Los valores más frecuentes se encuentran en el rango de tiempos mínimos, mientras que los casos en los que la duración es elevada son excepcionales y representan a un grupo reducido de usuarios. Este comportamiento refuerza lo observado previamente para la variable Informational: la mayoría de los usuarios no considera necesario profundizar en la información adicional disponible en el sitio, lo que sugiere un uso muy limitado de este tipo de páginas dentro del flujo de navegación típico.

Las duraciones asociadas a páginas administrativas también presentan una distribución sesgada hacia valores bajos, aunque con una mayor variabilidad en comparación con las páginas informativas. Si bien son muchos los usuarios que pasan poco tiempo en estas secciones, se observan casos aislados con tiempos significativamente superiores, lo que refleja que ciertos usuarios pueden estar realizando tareas más complejas o prolongadas, como gestiones de cuentas, lectura de términos o configuraciones específicas. Esta distribución heterogénea indica que el rol de las páginas administrativas es menos frecuente pero más relevante en los casos en que se utilizan, extendiendo la sesión cuando el usuario tiene una necesidad puntual.

La duración en páginas de productos exhibe valores considerablemente más elevados y dispersos en comparación con las duraciones de los otros tipos de páginas. Si bien la distribución también se encuentra sesgada a la derecha, su rango es mucho más amplio, con usuarios que pasan tiempos sustanciales navegando productos, evaluando alternativas o analizando información relevante para una posible compra. La concentración principal de valores se da en tiempos moderados, pero existen casos que registran duraciones muy altas, lo cual coincide con un comportamiento de exploración intensiva. Esto confirma que las páginas de productos no solo son las más visitadas, sino también aquellas donde los usuarios invierten la mayor parte de su tiempo, convirtiéndolas en el componente central del proceso de decisión dentro del sitio.

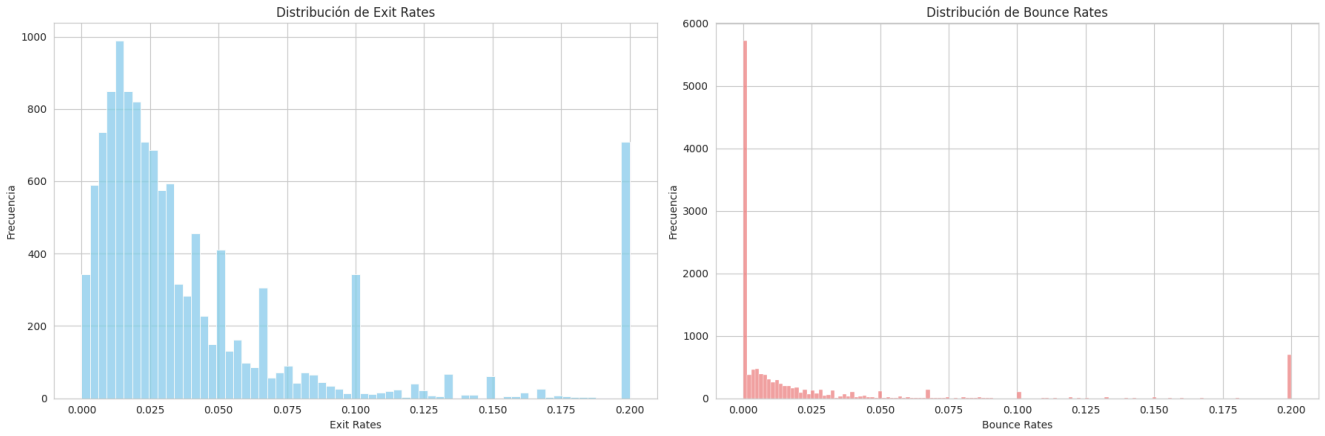
Luego encontramos “Page Value”, variable cuantitativa continua. Representa el valor promedio de la sesión completa, calculado a partir de la importancia histórica que tuvieron las páginas visitadas para concretar una venta.



La distribución de la variable PageValues muestra un patrón fuertemente concentrado en el valor cero, con la gran mayoría de las sesiones sin aportes significativos en términos de valor asociado a la conversión. Esto se debe a que sólo ciertas combinaciones de páginas visitadas o secuencias de navegación incrementan el valor histórico de una sesión de acuerdo con su relevancia para concretar ventas pasadas. Sin embargo, aunque los valores positivos son menos frecuentes, cuando aparecen suelen presentar una dispersión considerable, reflejando sesiones en las que el usuario transita por páginas con alto impacto en el proceso de compra. Esta distribución altamente asimétrica sugiere que la mayoría de los usuarios no llega a recorrer páginas críticas para la conversión, mientras que una minoría, asociada probablemente a usuarios más cercanos a la compra, genera valores notablemente mayores. Esto convierte a PageValues en una variable especialmente útil para identificar comportamientos con intención transaccional o etapas avanzadas del embudo de conversión.

La columna Bounce Rate (variable cuantitativa continua) se refiere al porcentaje de sesiones que finalizaron tras la visita a una sola página, es decir, cuando el usuario ingresó al sitio y lo abandonó sin realizar ninguna otra interacción (como hacer clic en enlaces o navegar hacia otra sección).

De forma similar, Exit Rate representa el porcentaje de últimas vistas para cada página, reflejando qué tan frecuentemente una sesión finaliza en una determinada sección del sitio.

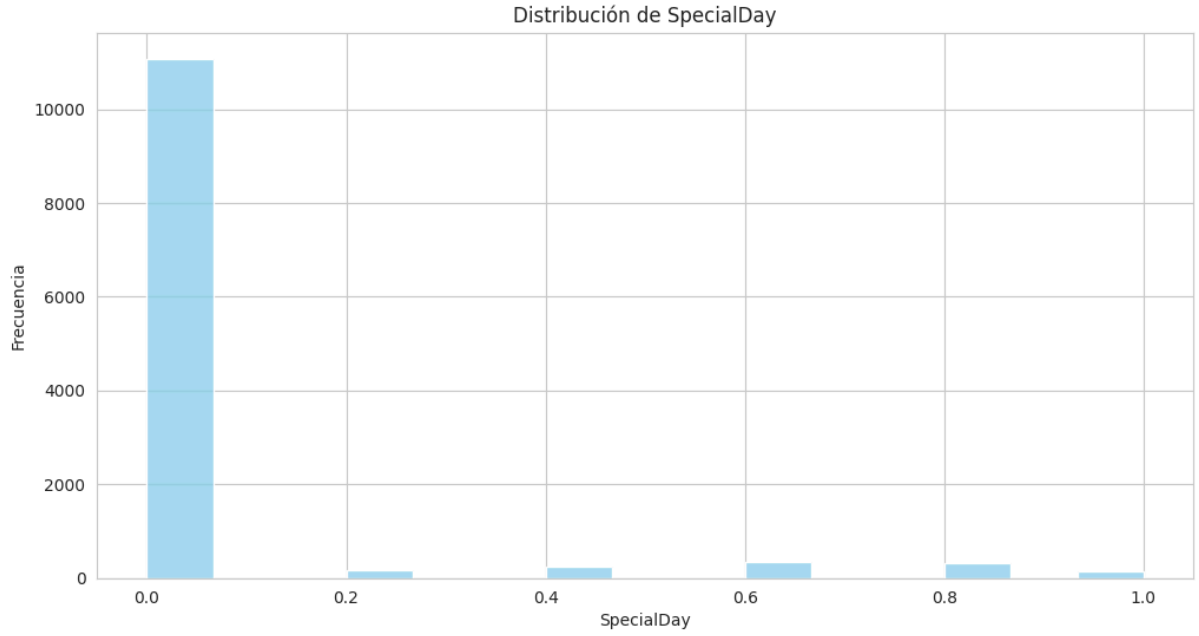


La distribución de Exit Rates muestra una fuerte concentración en valores muy bajos, lo cual indica que la mayoría de las páginas del sitio no suelen ser el punto final de la sesión en la mayoría de los casos. La asimetría marcada y la presencia de una larga cola derecha revelan que solo un conjunto reducido de páginas acumula porcentajes más altos de salidas. Esto sugiere que ciertas secciones funcionan como “puntos de fuga” dentro del sitio, ya sea porque el usuario considera que ha completado su objetivo o porque encuentra barreras que lo llevan a abandonar la navegación. La

variabilidad presente en estos valores hace de Exit Rates un indicador útil para identificar páginas que podrían presentar problemas de fricción o falta de continuidad dentro del flujo de navegación.

El histograma de Bounce Rates refleja un patrón similar, con la mayor parte de las observaciones concentradas en valores cercanos a cero. Esto implica que, en la mayoría de las sesiones, los usuarios interactúan al menos con más de una página antes de abandonar el sitio. Al igual que en ExitRates, la distribución es altamente asimétrica y presenta un grupo minoritario de páginas con tasas de rebote más elevadas. Las páginas con Bounce Rates altos representan potenciales puntos de desinterés o desalineación entre lo que el usuario esperaba encontrar y lo que realmente ofrece la página de aterrizaje. Debido a esta característica, la variable se convierte en una métrica clave para evaluar la eficacia de las páginas de entrada y para detectar problemas de relevancia, experiencia de usuario o carga de contenido.

Luego si observamos la columna Special Day (variable cuantitativa continua) representa la cercanía temporal entre la fecha de visita del usuario y un día especial, como el Día de la Madre o San Valentín. Su valor varía entre 0 y 1, donde 1 indica que la visita ocurrió el mismo día especial y valores intermedios reflejan que la sesión tuvo lugar en los días previos.



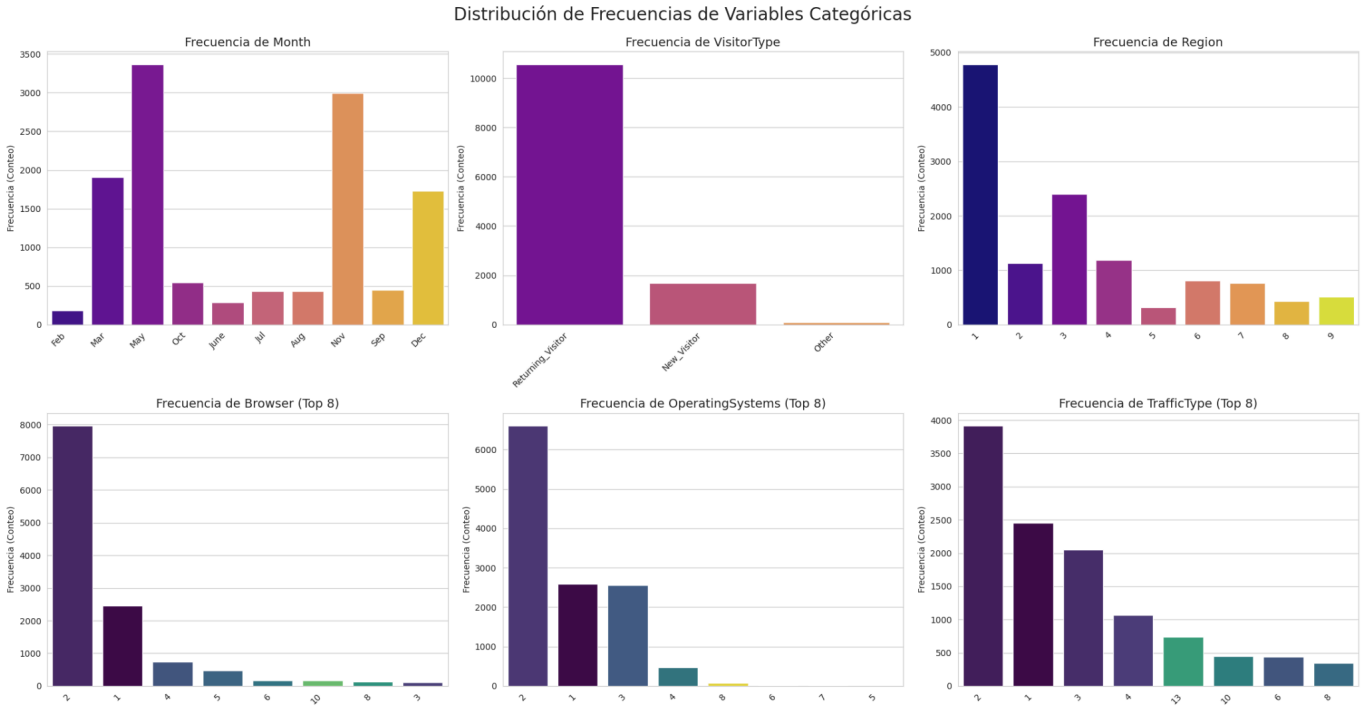
La distribución de SpecialDay se encuentra fuertemente concentrada en el valor cero, lo que indica que la mayor parte de las sesiones no ocurren en fechas cercanas a un día especial. Sin embargo, al observar la escala de valores, se identifica un aspecto importante: aunque la variable está definida como cuantitativa continua entre 0 y 1, en la práctica los datos no toman cualquier valor dentro de ese rango. Por el contrario, SpecialDay se presenta discretizada en incrementos regulares de 0,2 (0, 0,2, 0,4, 0,6, 0,8 y 1). Esto indica que el nivel de cercanía respecto a un día especial no se mide de forma continua, sino en categorías predefinidas que representan distintos intervalos de anticipación.

Los valores distintos de cero aparecen en una proporción muy baja, lo cual refleja que sólo una minoría de usuarios visita el sitio en los días previos o durante eventos especiales. La presencia escalonada de estos valores sugiere que el sitio define ciertos umbrales temporales —por ejemplo, 5 días antes, 4 días antes, etc.— para establecer la cercanía al evento. Este comportamiento confirma que las fechas especiales, aunque relevantes desde el punto de vista comercial, representan un volumen reducido dentro del conjunto total de sesiones. En conjunto, la variable SpecialDay evidencia tanto la baja frecuencia de navegación en días estratégicos como un diseño de medición que clasifica la cercanía temporal en categorías discretas y no en valores continuos.

En cuanto a las variables categóricas, Month es cualitativa ordinal, ya que representa los meses del año con un orden natural. Operating Systems es una variable categórica nominal, que indica el sistema operativo utilizado por el visitante; aunque está codificada numéricamente, su naturaleza es categórica. Del mismo modo, Browser y Región son variables cualitativas nominales.

Traffic Type describe la fuente de tráfico que llevó al usuario al sitio (por ejemplo, directo, por buscadores o referidos). tratarla como variable de clasificación.

Visitor Type distingue distingue entre nuevos visitantes (*New_Visitor*) y visitantes recurrentes (*Returning_Visitor*), siendo también una variable categórica nominal.



La variable Month muestra una distribución irregular, con ciertos meses registrando un volumen notablemente mayor de sesiones que otros. Esto indica que la actividad del sitio no es homogénea a lo largo del año, lo cual puede estar relacionado con estacionalidad, campañas comerciales específicas, aumentos temporales en el tráfico o cambios en el comportamiento de los compradores. Meses como noviembre o diciembre suelen presentar picos debido a eventos como Black Friday o compras navideñas, mientras que otros meses exhiben menor actividad. Esta distribución refuerza la importancia de comprender los patrones temporales para optimizar promociones y estrategias de marketing.

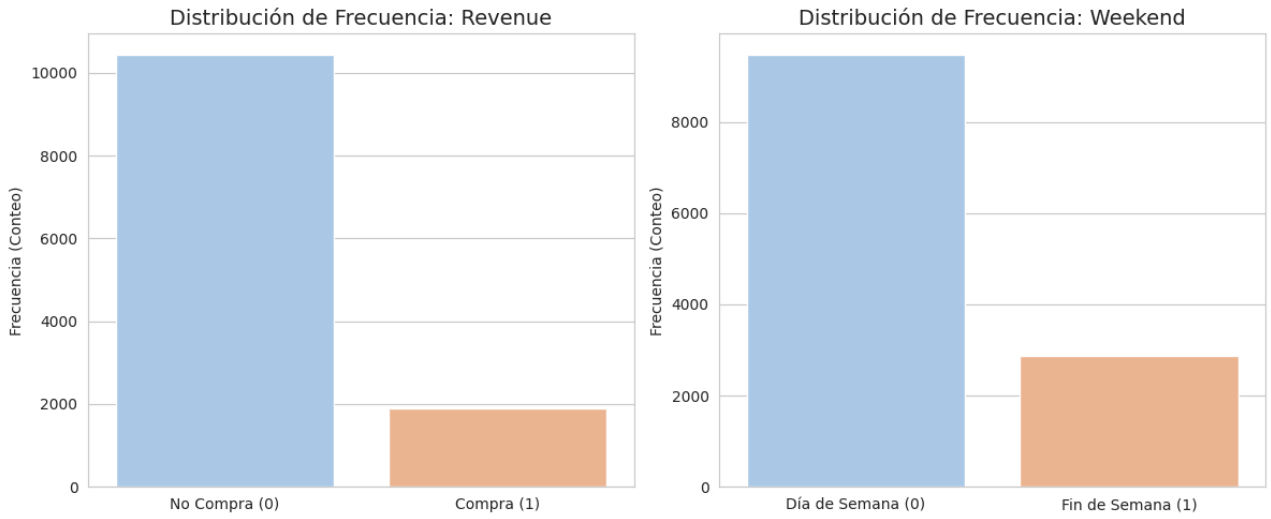
En VisitorType, se observa que los visitantes recurrentes representan la mayor proporción de sesiones, mientras que los nuevos visitantes conforman un grupo significativamente menor. Esta predominancia de usuarios recurrentes puede interpretarse como una señal positiva de fidelidad o hábito de uso del sitio, ya que indica que un número importante de visitantes regresa para explorar productos o realizar compras. Sin embargo, también sugiere que la captación de nuevos usuarios podría ser un área potencial de mejora desde el punto de vista comercial.

La variable Región exhibe una distribución marcada por unas pocas regiones dominantes, que concentran la mayor parte del tráfico, mientras que otras aparecen con frecuencias considerablemente menores. Esto puede deberse tanto a factores demográficos como al alcance geográfico del sitio o a diferencias en hábitos de compra entre regiones. Las variaciones observadas sugieren que algunas zonas presentan una mayor afinidad con la plataforma, lo cual podría orientar estrategias de segmentación o campañas específicas.

En la distribución de Browser, se observa que unos pocos navegadores concentran la mayor parte del uso. La grilla se centra en los ocho más frecuentes, y aun dentro de este subconjunto existen diferencias marcadas. Navegadores populares como Chrome o Firefox tienden a dominar el tráfico, mientras que el resto aparece con menor representación. Esta información es relevante para decisiones técnicas del sitio, ya que indica en qué entornos de navegación conviene optimizar funcionalidades y pruebas de compatibilidad.

La variable OperatingSystems presenta un patrón similar, con uno o dos sistemas operativos que concentran la mayoría de las sesiones. Esto refleja la predominancia de ciertos dispositivos o plataformas entre los visitantes, lo cual también tiene implicancias técnicas y comerciales. Un sistema operativo claramente dominante suele estar asociado a smartphones o sistemas de escritorio mayoritarios, mientras que otros aparecen en proporciones menores pero no despreciables.

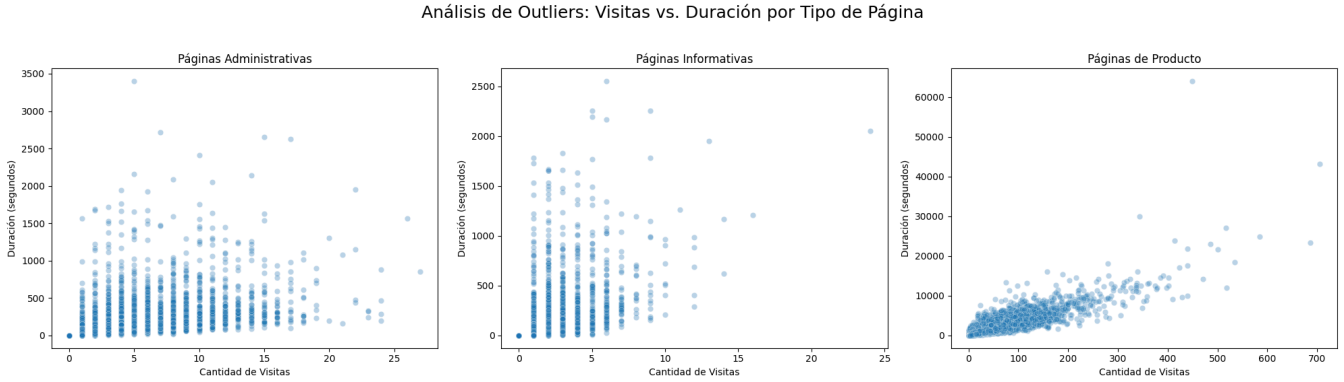
Finalmente, *Weekend* y *Revenue* son variables dicotómicas (valor si o no/verdadero o falso). La primera indica si la compra se realizó durante un fin de semana, y la segunda si esta visita generó ingresos para la página.



La variable Weekend muestra una distribución claramente desequilibrada, con una mayor cantidad de sesiones realizadas durante días de semana en comparación con los fines de semana. Esto indica que el tráfico del sitio se concentra mayormente en jornadas laborales, lo cual podría estar asociado a hábitos de navegación vinculados a horarios de oficina, búsqueda de productos durante la semana o simplemente mayor disponibilidad de usuarios en días hábiles. La menor proporción de sesiones durante el fin de semana sugiere que este periodo, si bien relevante, no constituye el principal motor de visitas para el sitio.

Por otro lado, la variable Revenue presenta una proporción significativamente baja de sesiones que finalizan en una transacción, lo cual es consistente con patrones típicos de navegación en sitios de comercio electrónico, donde solo una fracción minoritaria de visitas culmina efectivamente en una compra. La marcada asimetría observada pone en evidencia que la mayoría de los usuarios utiliza el sitio para explorar productos, comparar opciones o informarse, pero no necesariamente para concretar una transacción en la misma sesión.

Análisis de outliers:



El análisis de dispersión que relaciona la cantidad de visitas con la duración total por tipo de página (Administrativas, Informativas y de Producto) evidencia que los registros con valores elevados en ambos ejes no constituyen outliers anómalos ni errores de medición.

En los tres gráficos se observa un patrón consistente: a medida que aumenta la cantidad de visitas dentro de una sesión, también lo hace la duración total asociada. Si bien la dispersión es amplia, la tendencia general indica un incremento conjunto de ambas variables, lo cual es esperable desde el punto de vista del comportamiento del usuario.

Los puntos más extremos —como duraciones superiores a 60.000 segundos en páginas de producto— al considerar métricas normalizadas como el tiempo promedio por visita (Duración / Visitas), estos casos dejan de situarse en posiciones extremas y se alinean con el comportamiento del resto de la muestra, reforzando la interpretación de que se trata de sesiones de alto nivel de engagement.

En el resto de las variables tampoco se identificaron valores inverosímiles o incompatibles con la lógica del dominio, reafirmando la integridad del dataset.

En síntesis, los valores aparentemente outliers responden a patrones coherentes de uso intensivo y no presentan indicios de error. En consecuencia, no deben ser removidos del análisis, ya que forman parte legítima de la variabilidad natural del comportamiento de navegación de los usuarios.

Limpieza:

El proceso de limpieza del dataset fue relativamente sencillo debido a la buena calidad general de los datos. En primer lugar, se verificó la presencia de valores nulos en todas las variables y se constató que no existían registros incompletos, por lo que no fue necesario realizar imputaciones ni eliminar instancias por ausencia de información. Esto constituye un punto positivo, dado que garantiza que el análisis posterior no se vea afectado por técnicas de completado artificial.

Se procedió a eliminar registros duplicados, lo cual permitió asegurar que cada observación representa una sesión única. Esta depuración evita el sesgo que podría producir la repetición accidental de sesiones en los análisis estadísticos y en el entrenamiento posterior de modelos.

Para clarificar el análisis exploratorio (EDA) y permitir comparaciones más relevantes, se realizó un agrupamiento estratégico de categorías. En variables como OperatingSystems,

Browser y TrafficType, se observó una gran cantidad de modalidades con frecuencias muy bajas (inferiores al 1% del total).

Mantener estas categorías "raras" separadas complejiza innecesariamente la interpretación de los gráficos y las métricas. Más importante aún, impide realizar comparaciones justas y estadísticamente válidas (ej., comparar la mediana de un grupo de N=5000 contra uno de N=3 no tiene sentido).

Por ello, se agruparon en una categoría uniforme ("otros"). Esta estrategia nos permite enfocar el análisis en las categorías de alta frecuencia y, a su vez, comparar su comportamiento contra el grupo agregado de "otros", generando conclusiones más robustas.

Adicionalmente, esta simplificación es una práctica recomendada que también mejora la estabilidad de futuros modelos predictivos. Esto lo logra al reducir el ruido—es decir, se evita que el modelo aprenda patrones falsos de categorías con muy pocos datos—y al mitigar la sparsity (fragmentación), que es el problema que surge cuando los datos están tan dispersos en múltiples categorías pequeñas que se vuelve imposible encontrar patrones significativos.

Dado que el mes posee un carácter temporal y ordinal, se realizó un mapeo manual a valores numéricos respetando la secuencia cronológica. Esta decisión facilita su utilización en modelos que requieren inputs numéricos y evita que se la trate como una categoría nominal sin orden implícito.

Las variables Weekend y Revenue, originalmente booleanas, se codificaron como enteros (0 y 1). Esto simplifica su manipulación en análisis estadísticos y modelos de machine learning.

3. Hipótesis planteadas y resolución

En esta sección se presentan las hipótesis desarrolladas a partir de las observaciones obtenidas durante el análisis exploratorio. Cada hipótesis busca evaluar cómo distintos factores del comportamiento del usuario influyen en la probabilidad de compra (Revenue) dentro del sitio web.

Los usuarios que realizan compras demuestran un mayor engagement, medido como tasa de tiempo por visita (Tiempo_Total) más alto, en comparación con los que no compran.

La hipótesis busca establecer una relación entre el tiempo que pasa el usuario y si termina comprando o no. La "duración total" bruta es engañosa, ya que está sesgada por la cantidad de páginas vistas. Por ello, se construyó la métrica *Tiempo_Total*, que mide la calidad del engagement. Esta se calcula obteniendo primero la tasa de duración promedio por visita (Duración/Visitas) para cada categoría (Admin, Info, Producto).

Primero se creó la nueva variable “Tiempo_Total”. Para ello, se calculó la tasa de tiempo promedio por cada tipo de página: $Tasa\ de\ Tiempo\ = \frac{Duración\ del\ Tipo\ de\ Página}{Número\ de\ Visitas\ del\ Tipo\ de\ Página.}$

El uso de *np.where* garantiza que si no hubo visitas en una categoría, la tasa sea 0, evitando así errores de división por cero. El *Tiempo_Total* es la suma de estas tres tasas, resultando en una métrica de engagement no sesgada por el volumen de clics.

Análisis Estadístico de los Datos

Se deben comparar las distribuciones de *Tiempo_Total* entre los dos grupos: *Revenue*=1 (compran) y *Revenue*=0 (no compran).

Resultados de Tests (Normalidad y Homocedasticidad)

Test	Variable/Grupos	Estadístico	p-valor	Conclusión (al alpha=0.05)
Shapiro-Wilk	Revenue=1	0.638	0.000	**Rechaza Normalidad**
Shapiro-Wilk	Revenue=0	0.520	0.000	**Rechaza Normalidad**
Levene	Revenue=1 vs Revenue=0	4.019	0.045	**Rechaza Homocedasticidad** (las varianzas son diferentes)

Detalle de Normalidad (Shapiro-Wilk): Los p-valores de 0.000 para ambos grupos son significativamente menores que el nivel de significancia alfa=0.05. Esto implica que la distribución de la métrica *Tiempo_Total* no sigue una distribución normal en ninguno de los grupos.

Detalle de Homocedasticidad (Levene): El p-valor de 0.045 es menor que alpha=0.05, lo que indica que las varianzas de *Tiempo_Total* son significativamente diferentes entre los grupos.

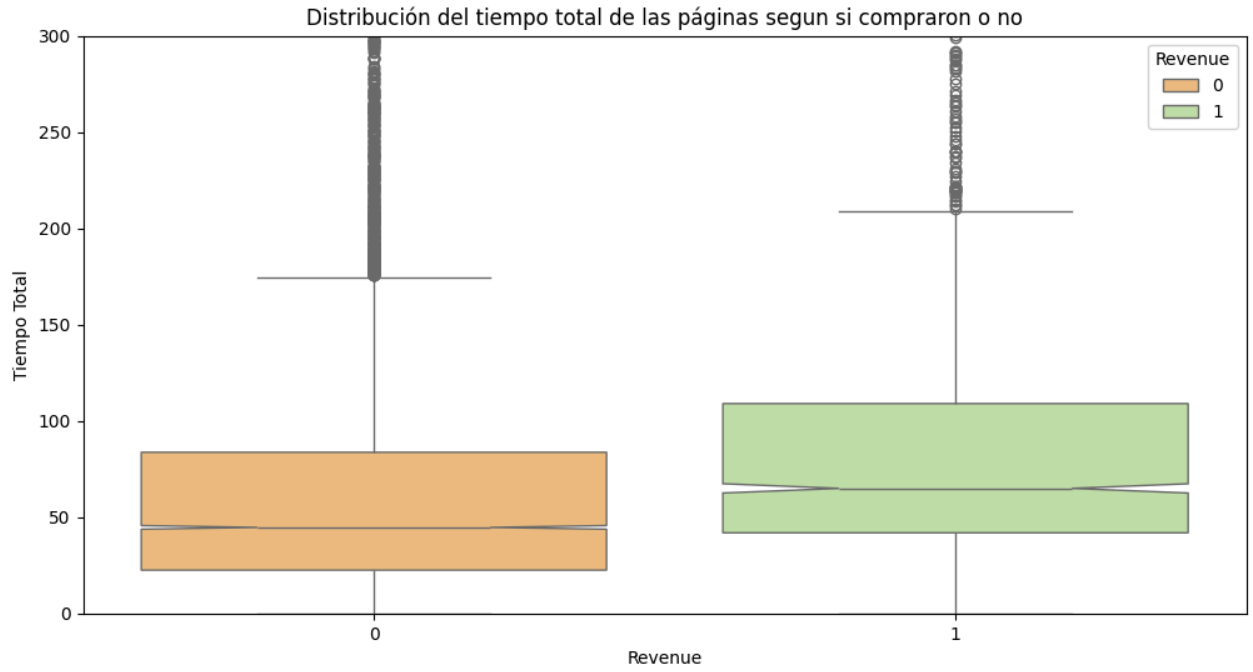
Tests No Paramétricos

Test	Estadístico	p-valor	Conclusión (al alpha=0.05)
Kruskal-Wallis	379.360	0.000	Se Rechaza la Hipótesis Nula
Mann-Whitney U	12576558.500	0.000	Se Rechaza la Hipótesis Nula

El p-valor de 0.000 es mucho menor que $\alpha = 0.05$, por ende se rechaza la Hipótesis Nula (H0: Las medianas de *Tiempo_Total* son iguales para Revenue=1 y Revenue=0). Por lo tanto, existe una diferencia estadísticamente significativa en la métrica *Tiempo_Total* entre los usuarios que compraron y los que no.

Análisis de los gráficos

El análisis del boxplot confirma la dirección de la diferencia encontrada.



Grupo *Revenue*=0 (No Compran - Naranja): La mediana (línea horizontal dentro de la caja) está aproximadamente en 45 unidades de tiempo.

Grupo *Revenue*=1 (Compran - Verde): La mediana está visiblemente más alta, aproximadamente en 65 unidades de tiempo.

La caja del grupo de compradores (*Revenue*=1) es más ancha y está posicionada más arriba en el eje Y que la del grupo de no conversión (*Revenue*=0). Esto indica que, en general, los usuarios que realizaron una compra pasaron significativamente más tiempo por visita promedio en las páginas que el grupo que no compró.

Las muescas (notch=True) representan el intervalo de confianza del 95% alrededor de la mediana. La falta de superposición de las muescas es una fuerte evidencia visual de que la diferencia entre las medianas es estadísticamente significativa.

Conclusión

Aceptamos la Hipótesis 1: Se encontró evidencia estadística significativa para afirmar que los usuarios que realizan compras (*Revenue* = 1) demuestran un mayor engagement, medido por la Tasa de Tiempo por Visita (*Tiempo_Total*), en comparación con los que no compran (*Revenue* = 0). El *Tiempo_Total* es un indicador fiable de la intención de compra. El simple hecho de "cliquear" no es tan importante como la "calidad" de ese tiempo de permanencia.

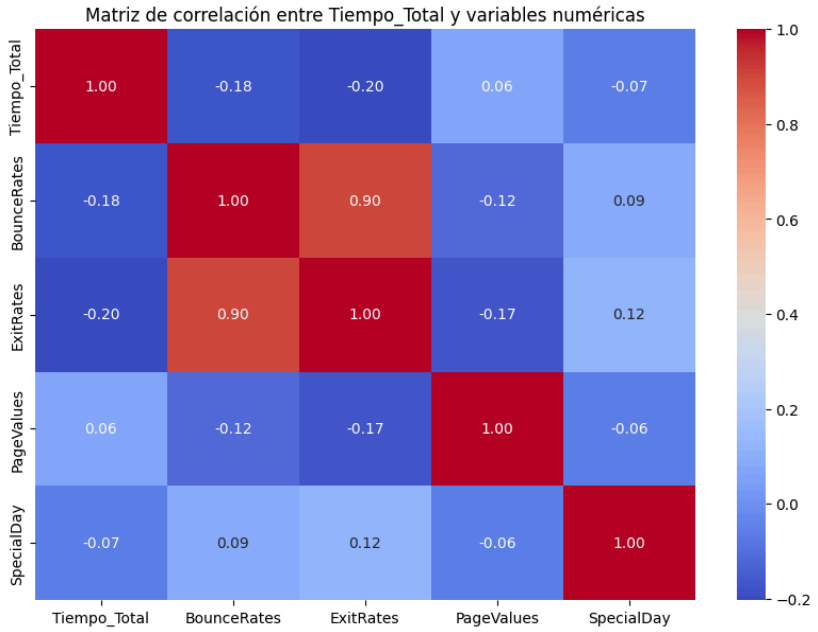
Existen factores, como Bounce Rates, ExitRates, PageValues y SpecialDay, que influyen significativamente en la tasa de tiempo por visita de los usuarios en el sitio

En este análisis, nuestra métrica *Tiempo_Total* deja de ser un predictor (variable independiente) y se convierte en nuestro objetivo (variable dependiente).

El objetivo es determinar si otras métricas clave — como las tasas de abandono (*BounceRates*, *ExitRates*), el valor de página (*PageValues*) y el contexto temporal (SpecialDay) — tienen poder explicativo sobre la variación del *Tiempo_Total*. Si podemos entender qué influye en el tiempo de visita, el negocio puede tomar acciones directas (ej. optimizar páginas con alto *ExitRate*) para mejorar el engagement y, como ya sabemos por la hipótesis 1, aumentar la probabilidad de compra.

Análisis de modelo lineal:

Se utilizó un mapa de calor (sns.heatmap) para visualizar las correlaciones de las variables numericas continuas elegidas con *Tiempo_Total,* y las correlaciones entre ellas.



Se evidencia que no existen correlaciones fuertes entre el Tiempo_Total y las demás variables, lo cual anticipa que un modelo lineal podría no tener gran capacidad explicativa. Sin embargo, este análisis sigue siendo valioso para comprender las interacciones entre variables.

Se destacan algunos puntos interesantes:

BounceRates y *ExitRates* presentan una correlación extremadamente alta (≈ 0.90), lo que indica que ambas métricas reflejan comportamientos muy similares de abandono del sitio. Por lo tanto, incluir ambas en un mismo modelo podría generar problemas de multicolinealidad y afectar la estabilidad de los coeficientes.

PageValues y *SpecialDay* muestran correlaciones débiles con el tiempo total, pero podrían tener un efecto combinado o no lineal que valga la pena explorar.

Las correlaciones negativas entre *Tiempo_Total* y las tasas de rebote/salida sugieren que a mayor interacción del usuario (más tiempo), menores son las probabilidades de abandono, lo cual es coherente con el comportamiento esperado.

En síntesis, aunque las correlaciones iniciales no son alentadoras para esperar un modelo lineal altamente predictivo, se procede con la regresión para evaluar si existen factores que impacten significativamente en el tiempo total de permanencia de los usuarios.

Al armar y ver los resultados del modelo lineal se concluye lo siguiente:

El valor de $R^2 = 0.035$ indica que el modelo explica solo el 3,5% de la variabilidad total del tiempo en página, lo que refleja un bajo poder explicativo.

Esto sugiere que el comportamiento del tiempo total no puede modelarse adecuadamente con una relación lineal simple entre estas variables: probablemente intervengan factores no lineales o interacciones complejas (como el tipo de visitante, la intención de compra o el tipo de página visitada).

A pesar del bajo R^2, el estadístico $F = 147.2$ con un p-valor < 0.001 indica que el modelo en su conjunto es estadísticamente significativo. Es decir, aunque la capacidad predictiva global sea limitada, al menos una de las variables independientes tiene un efecto lineal significativo sobre *Tiempo_Total*.

Los resultados del test Ómnibus y Jarque-Bera ($p \approx 0.000$) muestran que los residuos no siguen una distribución normal, confirmando que los supuestos clásicos de la regresión lineal no se cumplen completamente.

Esto implica que, si bien el modelo puede describir tendencias generales, no es óptimo para realizar inferencias precisas sobre la población ni para usarlo como modelo predictivo robusto.

Exploración No Lineal: Reducción de Dimensionalidad (t-SNE y UMAP)

Para buscar estructuras no lineales, se aplicaron las técnicas de reducción de dimensionalidad t-SNE y UMAP a las variables predictoras, y se colorearon los gráficos según el valor de Tiempo_Total. Esto permite visualizar si las variables predictoras generan agrupamientos que se correspondan con altos o bajos tiempos de permanencia.

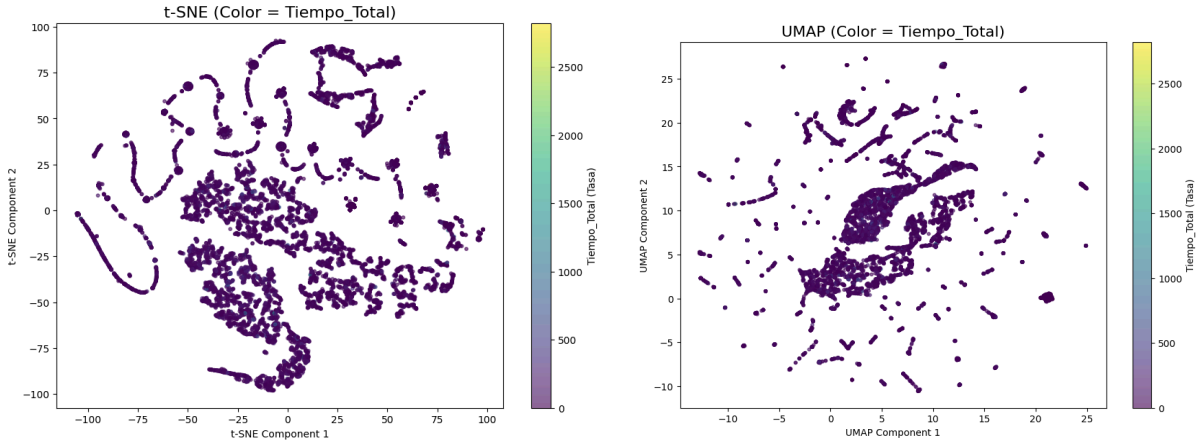
Análisis de los Gráficos (t-SNE y UMAP):

Dado que el modelo lineal no mostró una capacidad explicativa satisfactoria sobre el *Tiempo_Total*, se decidió aplicar técnicas de reducción de dimensionalidad no lineales (t-SNE y UMAP) con el fin de explorar patrones ocultos en las variables del comportamiento del usuario.

En este caso, se utilizan únicamente variables numéricas continuas elegidas previamente para la hipótesis.

Se excluyen las variables categóricas (Browser, *OperatingSystem*, *Region*, *TrafficType*) ya que t-SNE y UMAP calculan distancias continuas y las categorías codificadas numéricamente podrían distorsionar los resultados.

Ambos métodos permiten reducir el espacio de características a dos dimensiones y visualizar cómo se agrupan las observaciones según las variables seleccionadas.



No se observan agrupamientos definidos ni gradientes claros en función del *Tiempo_Total*, lo que sugiere que las variables *BounceRates*, *ExitRates*, *PageValues* y SpecialDay no determinan de manera directa el tiempo promedio por visita.

Conclusión

Las visualizaciones refuerzan la conclusión del modelo lineal. Los factores *BounceRates*, *PageValues* y *SpecialDay* no generan patrones de vecindad que expliquen el tiempo total de navegación de los usuarios, ni en un espacio lineal ni en uno no lineal.

En síntesis, la Hipótesis 2 no se sostiene con el conjunto de variables de comportamiento consideradas. El tiempo que un usuario dedica al sitio parece depender de factores más complejos, posiblemente variables categóricas o interacciones entre ellas.

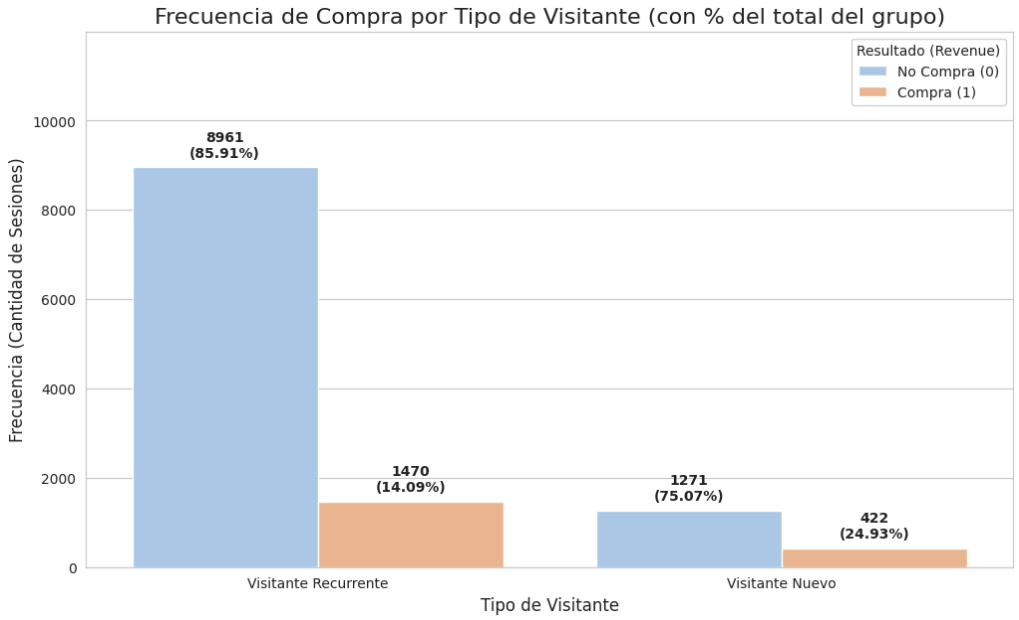
Aunque los Visitantes Recurrentes generan mayor volumen de compras, los Visitantes Nuevos demuestran una tasa de éxito (compran más fácil) estadísticamente superior

El objetivo de este análisis fue evaluar si el comportamiento de compra difiere significativamente entre los Visitantes Recurrentes y los Visitantes Nuevos, y determinar si la tasa de conversión (calidad) justifica una estrategia específica para cada grupo.

Antes de iniciar el análisis, se realizó una limpieza inicial del dataset para asegurar la relevancia de los datos, excluyendo a los visitantes clasificados como "other" debido a su bajo volumen y su poca representatividad para el análisis principal.

El primer paso exploratorio consistió en contar el total de sesiones por tipo de visitante:

Se generó un gráfico de barras agrupadas para visualizar la distribución de sesiones que terminaron en compra (Revenue = 1) versus las que no (Revenue = 0), calculando el porcentaje de conversión dentro de cada grupo de visitantes.



Interpretación del Gráfico

Al analizar el gráfico de barras, podemos definir la naturaleza de esta relación, la cual expone una dualidad clave (Volumen vs Calidad):

En Volumen Absoluto: Los Visitantes Recurrentes generan la gran mayoría de las transacciones (1,470 compras) en comparación con los Visitantes Nuevos (422 compras).

En Tasa de Conversión (Calidad): Los Visitantes Nuevos son significativamente más valiosos, demostrando una tasa de conversión (24.93%) que es casi el doble que la de los Visitantes Recurrentes (14.09%).

Test : Chi-Cuadrado

Para confirmar si la diferencia observada en las tasas de conversión (14.09% vs. 24.93%) es estadísticamente significativa y no resultado del azar, se aplicó un Test de Chi-cuadrado de Independencia. Este test es apropiado porque ambas variables (VisitorType y Revenue) son categóricas.

Hipótesis Nula (H0): El tipo de visitante es independiente del resultado de la compra (es decir, el tipo de visitante no influye en la tasa de compras).

Hipótesis Alternativa (H1): El tipo de visitante sí influye en el resultado de la compra (es decir, las proporciones de compra difieren entre grupos).

Resultados

Chi-cuadrado = 128.9831

p-valor = 0.0000

Grados de libertad = 1

Dado que el p-valor es 0.0000, un valor muy inferior al umbral de significancia estándar (α= 0.05), se rechaza la Hipótesis Nula (H0).Esto confirma estadísticamente que existe una asociación significativa entre el tipo de visitante y la decisión de compra.

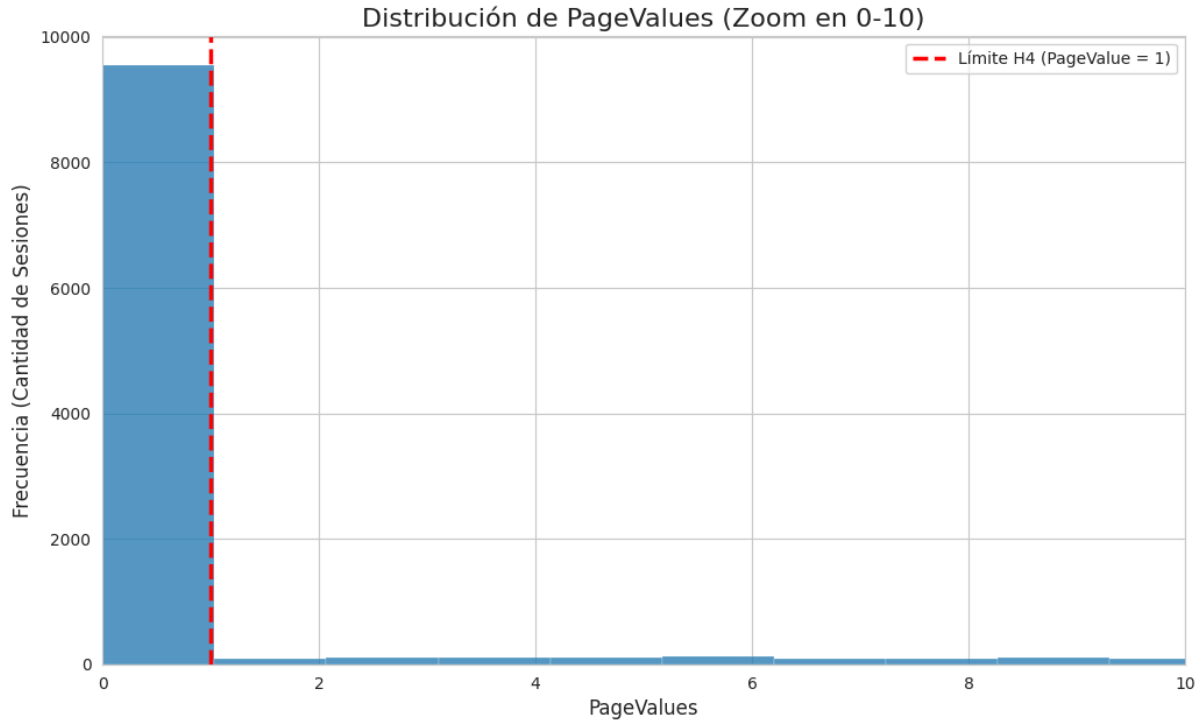
En conclusión la hipótesis se valida: mientras los Visitantes Recurrentes proveen el volumen de transacciones, los Visitantes Nuevos demuestran una tasa de conversión (calidad) estadísticamente superior. Esta diferencia debe ser considerada para optimizar las estrategias de marketing y diseño de experiencia de usuario (UX) dirigidas a cada segmento.

El 70% de los usuarios presentan un Page Value bajo (≤ 1), lo que indica una navegación meramente exploratoria, sin intenciones de compra.

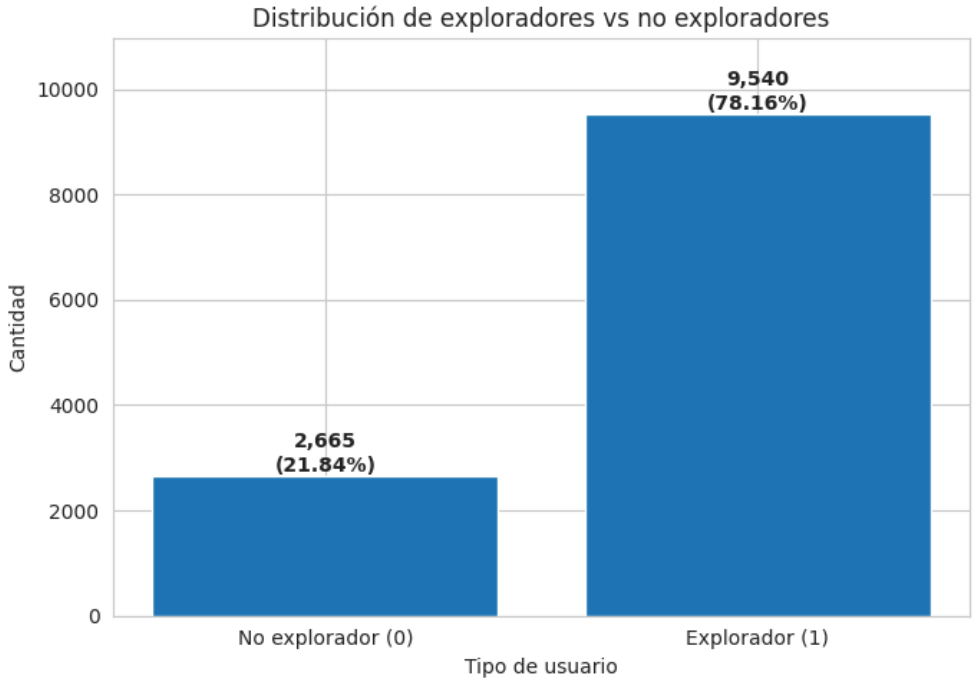
El Page Value es una métrica de Google Analytics que estima el valor promedio de una página para los ingresos. Un Page Value bajo (<1) se interpreta como una navegación de "bajo valor", es decir, la visita no contribuyó directamente a un ingreso significativo. La variable PageValue es, en teoría, uno de los indicadores más potentes de la intención de compra, ya que Google Analytics la diseña para medir cuánto contribuyó una página a una conversión.

Gráficos Utilizados

Se analizó la variable *PageValues* mediante gráficos de distribución y tests estadísticos, separando los datos en dos grupos: grupo1 (<= 1, Exploradores) y grupo2 (> 1, Alto valor).



El histograma recortado es el más efectivo. Permite ver claramente cómo la primera barra, que incluye el límite de 1, concentra la inmensa mayoría de los registros, mientras que la versión completa pierde detalle al estirar el eje X para incluir outliers. Acá usamos 'bins' altos para que las barras sean estrechas y ver mejor la diferencia ya que a simple vista en el histograma original, por una cuestión de la proporción no parecía que estuviéramos encontrando ninguna diferencia significativa. Se armó un gráfico de barras contabilizando la frecuencia de cada grupo para demostrar que no estamos engañando al público con los histogramas.



Formulación del Test

Se utiliza un test Z para una proporción (unilateral de cola derecha) con un nivel de significancia alfa = 0,05.

El Test Z se basa en el Teorema del Límite Central (TLC). Este teorema dice que si tu tamaño de muestra (N) es lo suficientemente grande (generalmente N > 30), la distribución de muestreo de la media (o proporción) será normal, sin importar cómo se distribuyen los datos originales.

Hipótesis Nula (H_0): La proporción real de usuarios con *PageValue* < 1 (exploradores) es mayor o igual al 70% (p >= 0,7)

Hipótesis Alternativa (H_1): La proporción real de usuarios con PageValue < 1 es menor al 70% (p < 0,7)

Resultados del Test Z

El código calcula la proporción observada (p sombrero) y el estadístico Z para contrastar con la proporción teórica p0=0.70.

Métrica	**Valor**
Proporción Observada (p sombrero)	0.7816 (78.16%)
Estadístico Z	21.8335
p-value	1.0000

Conclusión

La proporción observada de sesiones con PageValue ≤ 1 fue del 78.16%, un valor claramente superior al 70% planteado en la hipótesis nula. El estadístico obtenido (Z = 21.83) resultó marcadamente positivo, lo que indica que la proporción observada se

encuentra muy alejada de la región en la que se cumpliría la hipótesis alternativa ($p < 0.70$). En consecuencia, el p-valor asociado ($p = 1.0000$) señala que no existe evidencia estadística para afirmar que la proporción real sea inferior al 70%.

Con el test realizado no se rechaza la hipótesis nula ($p \geq 0.70$). Esto implica que no se encontró evidencia estadística para sostener que la proporción real de sesiones exploratorias sea inferior al 70%. Dado que la proporción observada en la muestra fue 78.16%, los datos son plenamente compatibles con que, en la población, al menos el 70% del tráfico corresponda a usuarios con navegación de bajo valor. Bajo un nivel de significancia del 5%, este resultado respalda la interpretación de que la mayoría del tráfico es efectivamente exploratorio.

La tasa de rebote varía significativamente según la región, mes y si la visita ocurrió en fin de semana o no

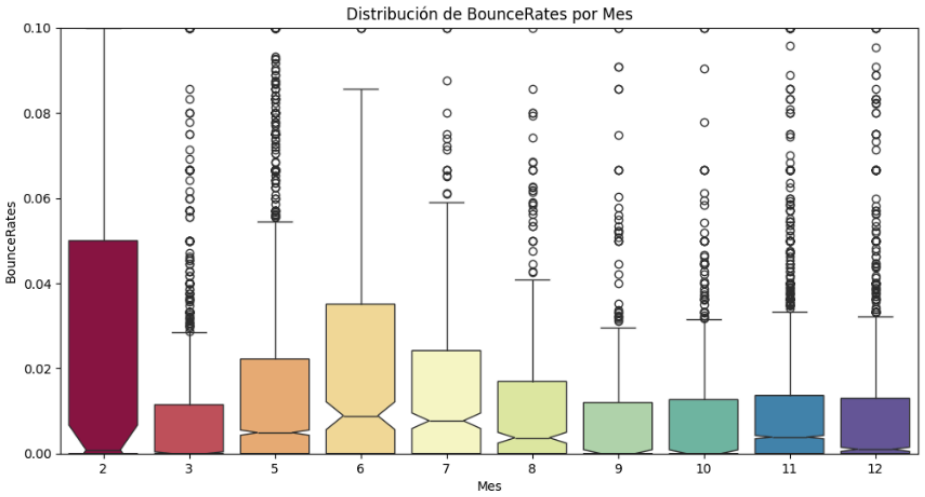
La tasa de rebote (BounceRate) refleja el nivel de interés o compromiso inicial del usuario con el sitio. Un rebote ocurre cuando el visitante abandona sin realizar más interacciones, por lo que variaciones en esta métrica pueden deberse a factores contextuales externos que afectan su comportamiento de navegación.

Los usuarios de distintas regiones pueden tener comportamientos heterogéneos por motivos culturales, económicos o técnicos (por ejemplo, diferencias en la conexión a internet, idioma o relevancia del contenido). Es razonable esperar que ciertas regiones presenten tasas de rebote más altas si el contenido no se ajusta a sus preferencias o necesidades.

La navegación web suele mostrar patrones estacionales: durante campañas de descuentos o eventos comerciales (como Black Friday o Navidad) los usuarios suelen permanecer más tiempo y rebotar menos. En cambio, en meses sin promociones, los rebotes pueden aumentar.

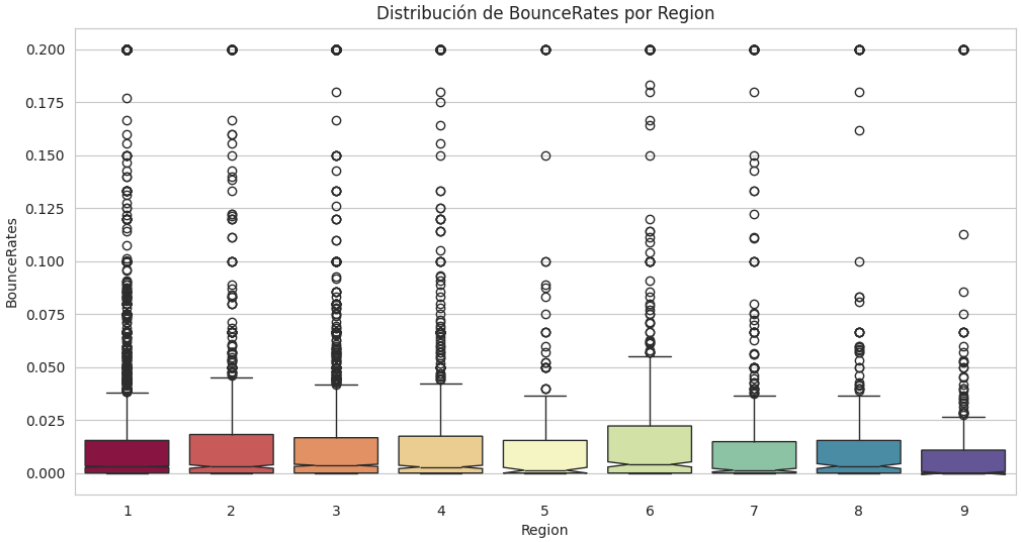
El comportamiento entre semana y fines de semana también suele diferir. Los usuarios de fin de semana suelen navegar de forma más casual o exploratoria, mientras que los de días hábiles pueden tener un propósito más concreto (por ejemplo, búsqueda de información específica o compra planificada).

Análisis de la visualización:



La tasa de rebote no es constante a lo largo del año. El Mes 2 (Febrero) destaca drásticamente sobre todos los demás, con una mediana (línea central) y una dispersión (tamaño de la caja) significativamente más altas.

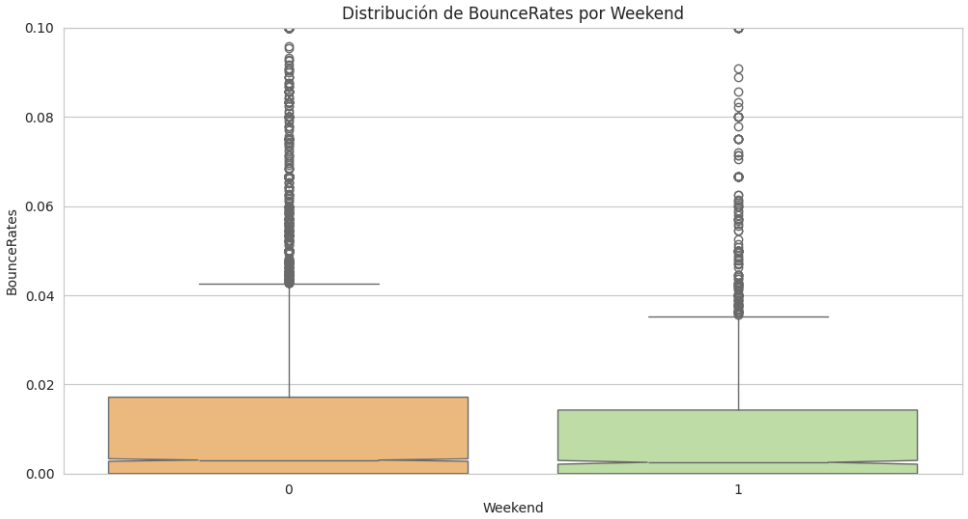
En los meses 3,9,10 y 12, que tienen distribuciones muy similares, se observa que la parte inferior a la mediana, está aplastada en 0, lo que indica que la mitad de las visitas realizadas esos meses, tuvieron un bounce rate nulo, lo cual es bueno, indica que el usuario hizo un acción relevante en la página que probablemente termine ejecutando una compra.



Analizando el ratio de saltos por región, vemos que las distribuciones se mantienen muy similares, las medianas no son muy diferentes unas de otras, inclusive los rangos máximos no difieren mucho entre se.

La región 6 pareciera tener un rango intercuartílico ligeramente más disperso que el resto, por ende un extremo máximo mayor. En esa región se ve una mezcla entre visitantes que quedan dentro de la página y los que no.

Algunas regiones (5, 7 y 9) presentan en el cajón de su gráfico de dispersión la parte inferior de la mediana está colapsada en cero, y esto significa que en esas regiones la mitad de sus visitas tuvieron acciones relevantes para la compra.



Además de que se observan distribuciones muy similares, el hallazgo más inmediato es que la línea de la mediana (Cuartil 2) para ambos grupos se sitúa un poco por encima de 0. Esto significa que, tanto en días de semana como en fines de semana, poco menos del 50% de todas las sesiones tuvieron una tasa de rebote muy baja.

La caja del grupo 0 (Día de Semana) es visiblemente más alta, con un tercer cuartil (Q3) situado aproximadamente en 0.017, mientras que la caja del grupo 1 (Fin de Semana) es más chata, con un Q3 más bajo (aprox. 0.013).

Tests realizados:

Test de normalidad:
En los 21 subgrupos analizados (10 por Mes, 9 por Región y 2 por Weekend), el p-valor resultante fue 0.0000. Dado que en todos los casos el p-valor es significativamente menor que el nivel de significancia estándar (alpha = 0.05), se rechaza de forma concluyente la hipótesis nula (H_0) de normalidad.

El incumplimiento del supuesto de normalidad descarta el uso de pruebas paramétricas (como ANOVA o el Test T de Student).

Para validar la Hipótesis 5 (comparar las BounceRates entre estos grupos), se deben utilizar los siguientes tests no paramétricos:

Test de Kruskal-Wallis: Para Month y Region (comparaciones de 3 o más grupos).

Test U de Mann-Whitney: Para Weekend (comparación de 2 grupos).

Conclusiones:

Los tests estadísticos confirman que las tres variables—Month, Region y Weekend—tienen una influencia estadísticamente significativa en la distribución de la BounceRate (tasa de rebote).

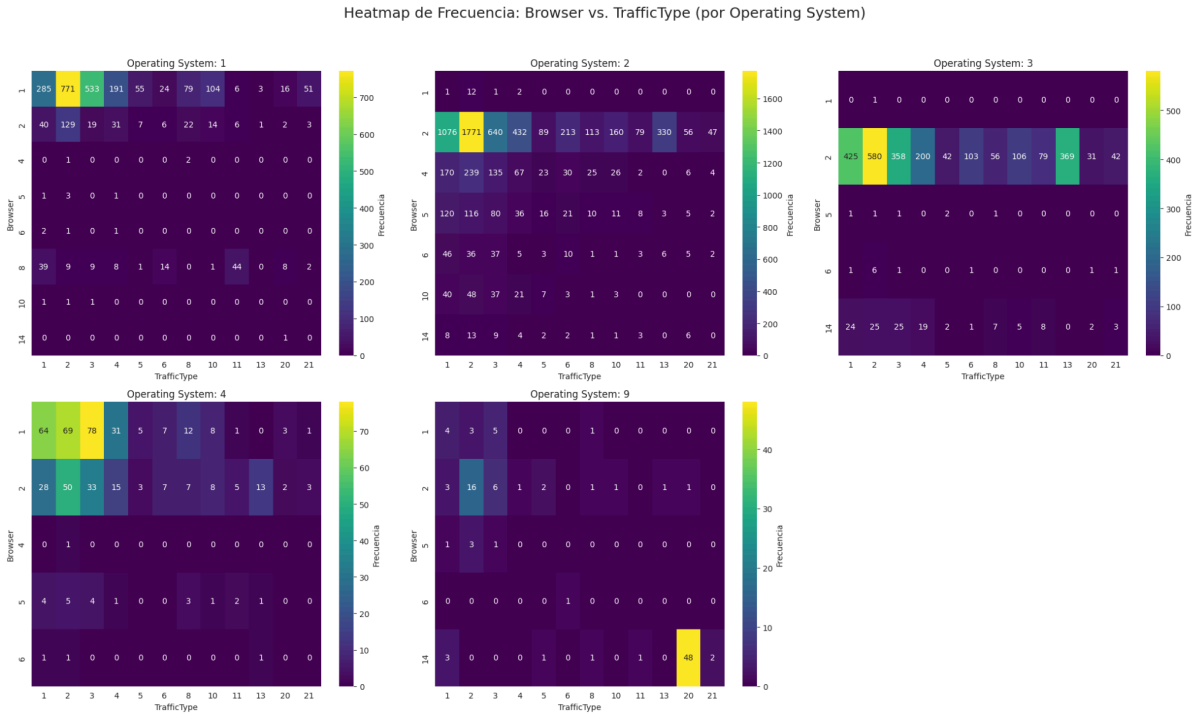
Tanto el Test de Kruskal-Wallis para Month ($p < 0.0001$) como para Region ($p < 0.0001$) arrojaron p-valores despreciables, indicando una fuerte relación. De igual manera, el Test U de Mann-Whitney para Weekend ($p = 0.0042$) resultó significativo, validando que existe una diferencia real en la tasa de rebote entre días de semana y fines de semana, aunque esta sea más sutil que la observada entre meses o regiones.

Según el Browser, tipo de tráfico y Operating System varía significativamente la tasa de abandono(Exit Rate)

Se postula que la tasa de abandono (Exit Rate) de una sesión es significativamente influenciada por la tecnología del usuario: Navegador (Browser), Tipo de Tráfico (TrafficType), y Sistema Operativo (Operating System). El objetivo es determinar si las diferencias en la distribución de la tasa de abandono observadas entre las distintas categorías tecnológicas son estadísticamente significativas.

La justificación de este análisis es puramente accionable: Si se descubre que un Browser,tipo de tráfico o Sistema Operativo específico tiene una tasa de abandono sistemáticamente más alta, esto es un fuerte indicador de problemas técnicos. Podría señalar bugs de compatibilidad, errores de renderizado, o scripts que no se ejecutan, los cuales frustran al usuario y deben ser corregidos por el equipo de desarrollo.

Analisis de la visualización



El heatmap visualiza la frecuencia de las sesiones de usuario en la intersección de las variables categóricas Browser y TrafficType. La intensidad del color, que varía de violeta (baja frecuencia) a amarillo brillante (alta frecuencia), permite identificar rápidamente los patrones de tráfico.

El análisis revela una distribución del tráfico extremadamente no uniforme y altamente concentrada. La inmensa mayoría de las sesiones se agrupan en solo unas pocas combinaciones dominantes:

Pico de Máxima Frecuencia: La intersección de Browser 2 y TrafficType 2 es el punto de mayor concentración y, con 3,192 sesiones, representa la combinación más común de tecnología y canal de adquisición.

Concentraciones Secundarias: A este pico principal le siguen otras tres concentraciones significativas, involucrando únicamente a los Browsers 1 y 2 y los TrafficTypes 1, 2 y 3:

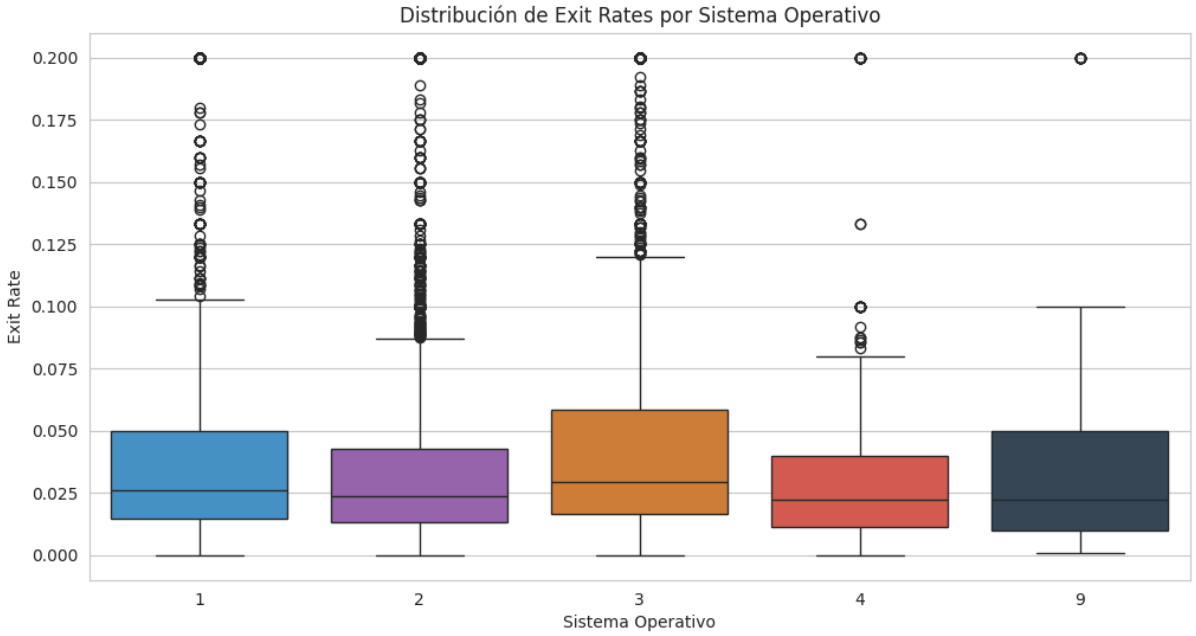
- Browser 2 / TrafficType 1 (1,391 sesiones)

- Browser 2 / TrafficType 3 (1,303 sesiones)
- Browser 1 / TrafficType 1 (1,017 sesiones)

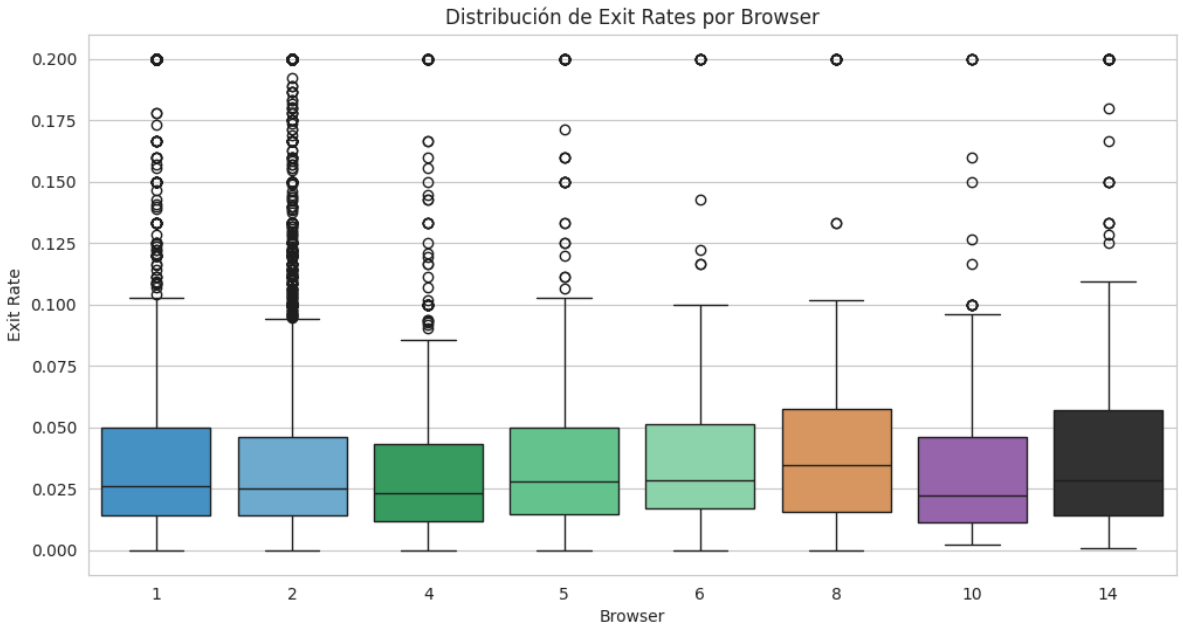
Fuera de estas cuatro combinaciones principales, la matriz es extremadamente dispersa. La vasta mayoría de las celdas se presentan en colores violeta oscuro, indicando frecuencias muy bajas (a menudo de un solo dígito).

Esta dispersión confirma que las categorías restantes de Navegador y Tipo de Tráfico constituyen una "larga cola" de comportamiento del usuario. Este hallazgo respalda la decisión metodológica de agrupar estas categorías minoritarias durante la limpieza y preprocesamiento de los datos para un análisis más enfocado.

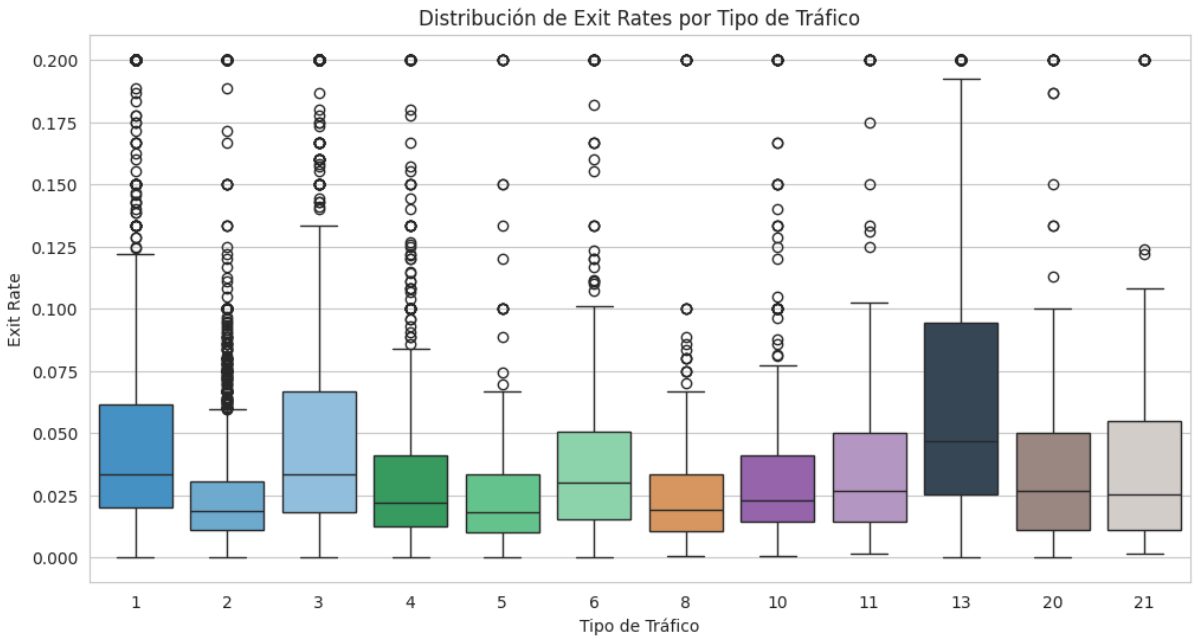
Las visualizaciones iniciales sugirieron la existencia de diferencias clave:



Exit Rate por Sistema Operativo : Los sistemas operativos 3 y 1 presentaron una mayor dispersión (cajas más altas) y medianas ligeramente superiores, lo que podría sugerir una experiencia de usuario menos optimizada en esos entornos.



Exit Rate por Navegador : Aunque las medianas se mantuvieron relativamente estables, los navegadores 8 y 14 mostraron una ligera elevación, sugiriendo potenciales problemas de compatibilidad o usabilidad específicos.



Exit Rate por Tipo de Tráfico : Se observaron las diferencias más notables, donde los tipos de tráfico 3, 13 y 20 mostraron medianas de abandono visiblemente más altas y una mayor dispersión. Esto indica que la fuente de adquisición es un factor de alto riesgo de abandono.

Evaluación de Supuestos (Tests de Normalidad)

Evaluación de Normalidad (Test de Shapiro-Wilk)
Como primer paso crucial, se utilizó el Test de Shapiro-Wilk para verificar la normalidad de la variable continua (Exit Rate) dentro de cada subgrupo categórico (Browser, Operating System y Traffic Type).

Conclusión del Test de Normalidad: Dado que en todos los subgrupos analizados el p-valor fue 0.0000 (significativamente menor que α=0.05), se rechazó de forma concluyente la hipótesis nula de normalidad. Esto implica que la distribución de la Exit Rate no sigue la curva normal en ninguna de las categorías, descartando el uso de pruebas paramétricas (como el ANOVA).

Test de hipótesis

Dado que el Test de Shapiro-Wilk confirmó que la variable Exit Rate no sigue una distribución normal en ninguno de los subgrupos, la herramienta estadística apropiada para comparar las medianas de una variable continua no normal entre múltiples grupos independientes fue el Test de Kruskal-Wallis. Se ejecutó un test separado para cada variable categórica.

Se realizaron tres tests de Kruskal-Wallis independientes, uno para cada variable categórica contra la ExitRate.

1. Kruskal-Wallis: ExitRate por OperatingSystem
   - Estadístico H = 136.2139
   - P-Valor = 0.0000

- Conclusión: Significativo ($p < 0.05$). El Sistema Operativo SÍ influye en la mediana de ExitRate.

2. Kruskal-Wallis: ExitRate por Browser
   - Estadístico H = 27.5454
   - P-Valor = 0.0003
   - Conclusión: Significativo ($p < 0.05$). El Navegador SÍ influye en la mediana de ExitRate.

3. Kruskal-Wallis: ExitRate por TrafficType
   - Estadístico H = 1351.4775
   - P-Valor = 0.0000
   - Conclusión: Significativo ($p < 0.05$). El Tipo de Tráfico SÍ influye en la mediana de ExitRate.

En los tres casos, el p-valor fue marcadamente inferior a 0.05 (nuestro nivel de significancia α). Esto significa que la probabilidad de observar las diferencias en las medianas de ExitRate entre los grupos (navegadores, OS, tipos de tráfico) si en realidad no existiera ninguna diferencia en la población (Hipótesis Nula, H0) es prácticamente nula. Por lo tanto, rechazamos firmemente la H0 para cada variable. Las diferencias observadas son estadísticamente significativas y no son producto del azar.

El Estadístico H del Test de Kruskal-Wallis actúa como una medida de la magnitud de las diferencias entre las medianas de los grupos. Una H más alta indica una mayor dispersión en las medianas de los grupos, lo que se traduce en una mayor influencia de esa variable categórica sobre la variable dependiente.

- TrafficType: H = 1351.5 (Máxima Influencia)
- OperatingSystem: H = 136.2 (Influencia Media)
- Browser: H = 27.5 (Menor Influencia)

La jerarquía de los estadísticos H revela que la variable que define cómo llega el usuario al sitio (TrafficType) es el predictor más fuerte y dominante del abandono, superando con creces la tecnología subyacente del usuario (OS y Browser). Esto sugiere que la calidad de la fuente de adquisición es clave para retener a los visitantes.

Conclusión

La hipótesis se valida. Se encontró evidencia estadística para afirmar que la mediana de la tasa de abandono (ExitRate) varía significativamente dependiendo del Sistema Operativo, el Navegador, y el Tipo de Tráfico del usuario.

Aunque las tres variables son significativas, el análisis del Estadístico H identifica al Tipo de Tráfico como el factor más crítico. Esto implica que las acciones de optimización más impactantes deben enfocarse en mejorar la calidad y relevancia del tráfico entrante, o adaptar la experiencia del sitio para manejar las altas tasas de abandono en canales de bajo rendimiento.