Trabajo Práctico Especial - Fundamentos de la Ciencia de Datos - 2025

Online Shoppers Purchasing Intention Dataset

Simón Abraham - Bautista Cisilino

Grupo N° 12

2do Cuatrimestre 2025

INTRODUCCIÓN

Breve contextualización : El e-commerce ha transformado la forma en la que los consumidores interactúan con las marcas, pero presenta un desafío permanente : la altísima tasa de 'abandono de carritos'. Existe un gran volumen de usuarios que navega, parece que muestra interés, pero no finaliza la transacción. Esta nueva forma de hacer compras ha generado cierto problema para las empresas, que son incapaces de distinguir entre los visitantes de su sitio web.

El espacio del problema que aborda este dataset principalmente es la predicción de la intención de compra en tiempo real. El objetivo es imitar la capacidad de un vendedor experimentado en una tienda física, utilizando 'lenguaje corporal digital' de un usuario que navega en el sitio web e identificar los momentos donde una oferta (entre otras opciones) podría ser la diferencia entre una visita fallida y una conversión exitosa.

Para explorar el problema, se utiliza el conjunto de datos "Online Shoppers Purchasing Intention". Una investigación sobre su origen revela que no son registros genéricos, sino el subproducto de un artículo académico del 2018.

El dataset se compone de 12,330 vectores/registros/muestras de características donde cada uno de estos registros representa una sesión de navegación **única** recopilada durante un año en un sitio web de comercio electrónico turco. La utilidad del conjunto de datos radica en el diseño: fue concebido para resolver un problema empresarial para predecir la compra en tiempo real.

ANÁLISIS EXPLORATORIO DE LOS DATOS

1. Descripción de las Variables

Para contextualizar el informe, se presenta un resumen de las variables contenidas en el dataset, clasificadas según su tipo :

- Administrative: Número total de páginas de tipo administrativo que el usuario visitó en la sesión. De tipo cuantitativa discreta.
- Administrative_Duration: Tiempo total (en segundos) que el usuario pasó en páginas de tipo administrativo durante la sesión. De tipo cuantitativa continua.
- Informational: Número total de páginas de tipo informativo que el usuario visitó en la sesión. De tipo cuantitativa discreta.
- Informational_Duration: Tiempo total (en segundos) que el usuario pasó en páginas de tipo informativo durante la sesión. De tipo cuantitativa continua.
- ProductRelated: Número total de páginas relacionadas con productos que el usuario visitó en la sesión. De tipo cuantitativa discreta.
- ProductRelated_Duration: Tiempo total (en segundos) que el usuario pasó en páginas relacionadas con productos durante la sesión. De tipo cuantitativa continua.
- BounceRates: Porcentaje de visitantes que ingresaron al sitio y lo abandonaron sin realizar ninguna acción adicional. De tipo cuantitativa continua.
- ExitRates: Porcentaje de visitas que finalizaron en una página específica, en relación con el total de visitas a esa página. De tipo cuantitativa continua.
- PageValues: Valor promedio (en unidades monetarias) aportado por cada página única visitada en una sesión que resultó en una conversión. De tipo cuantitativa continua.
- SpecialDay: Medida numérica que indica la proximidad de la visita a un día festivo (0 es máxima cercanía). De tipo cuantitativa continua.
- Month: Mes del año en que se realizó la visita (ej. "Feb", "Mar"). De tipo cualitativa ordinal.
- OperatingSystems: Código numérico que representa el sistema operativo del usuario. De tipo cualitativa nominal.
- Browser: Código numérico que representa el navegador web utilizado por el usuario. De tipo cualitativa nominal.
- Region: Código numérico que representa la región geográfica desde la cual el usuario realizó la visita. De tipo cualitativa nominal.

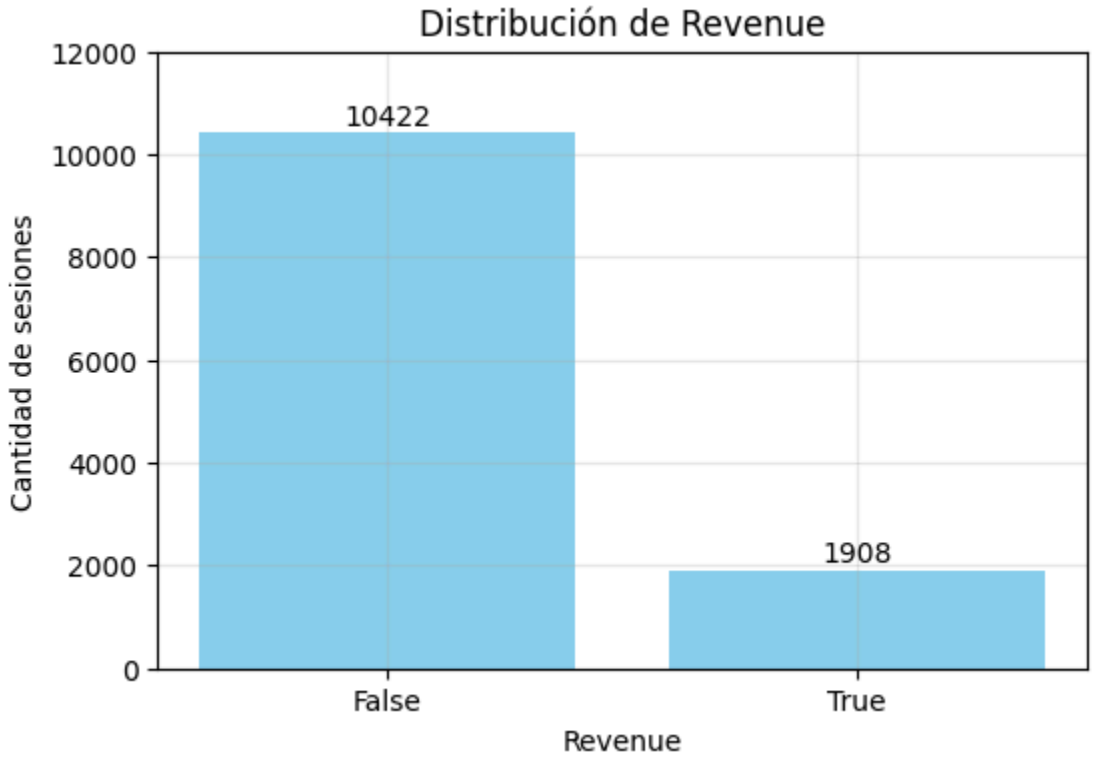
- TrafficType: Código numérico (1-20) que indica la fuente de tráfico por la cual el usuario llegó al sitio. De tipo cualitativa nominal.
- VisitorType: Categoría que clasifica al visitante (ej. "New Visitor", "Returning_Visitor"). De tipo cualitativa nominal.
- Weekend: Valor booleano (True/False) que indica si la sesión ocurrió durante un fin de semana. De tipo cualitativa nominal (binaria).
- Revenue: Valor booleano (True/False) que indica si la sesión finalizó en una compra (variable objetivo). De tipo cualitativa nominal (binaria).

2. Exploración univariada.

A continuación haremos un repaso de qué impresión tuvimos una vez que empezamos a navegar por los datos individualmente.

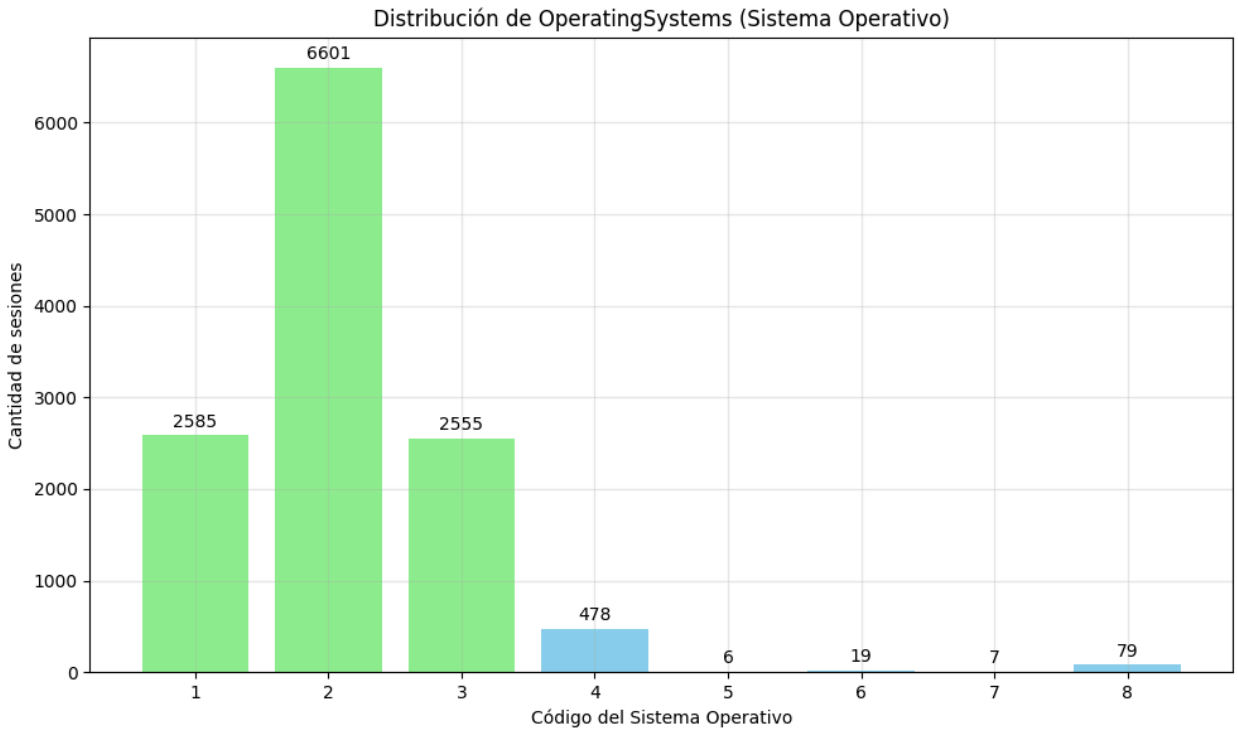
El dataset se compone de 18 variables de las cuales : 2 son booleanas, 7 son de tipo float, 7 son de tipo entero y 2 de tipo objeto.

Arrancamos por la variable que más importancia va a cargar en este dataset.Al fin y al cabo los interesados en esta información que nosotros brindamos se dejarán llevar por el resultado de si se compra o no (Revenue).

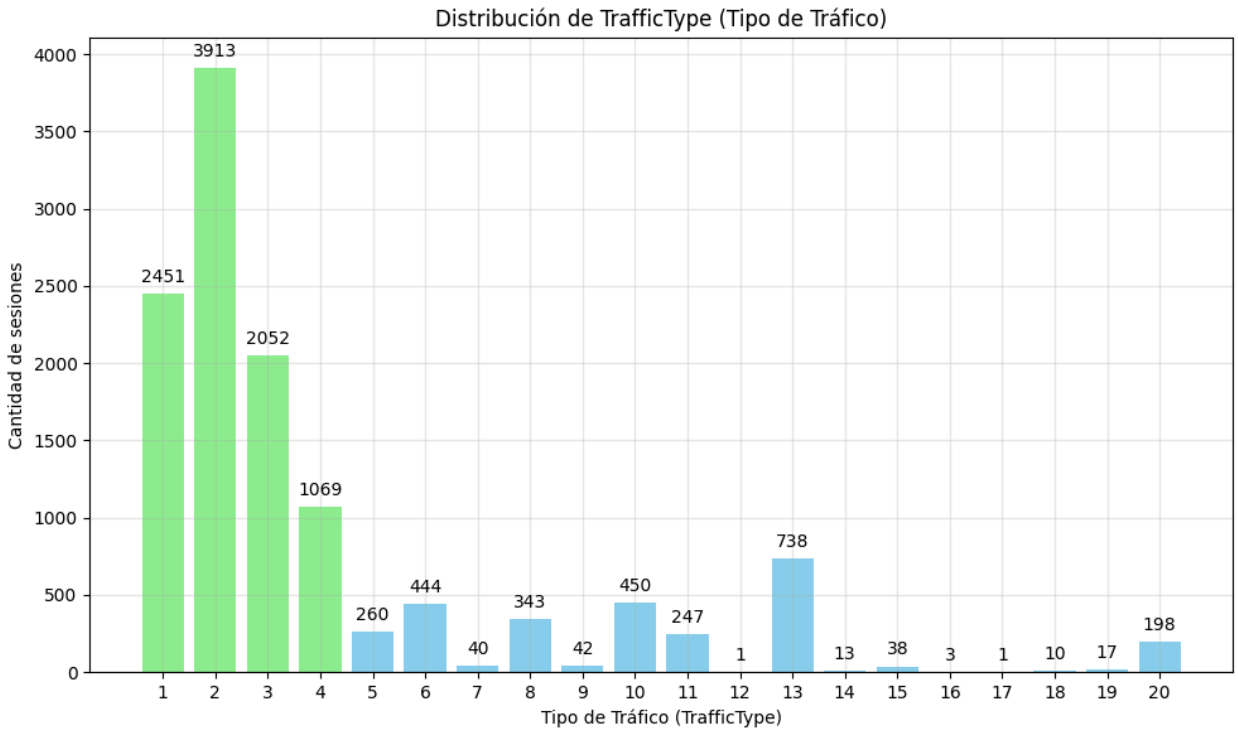


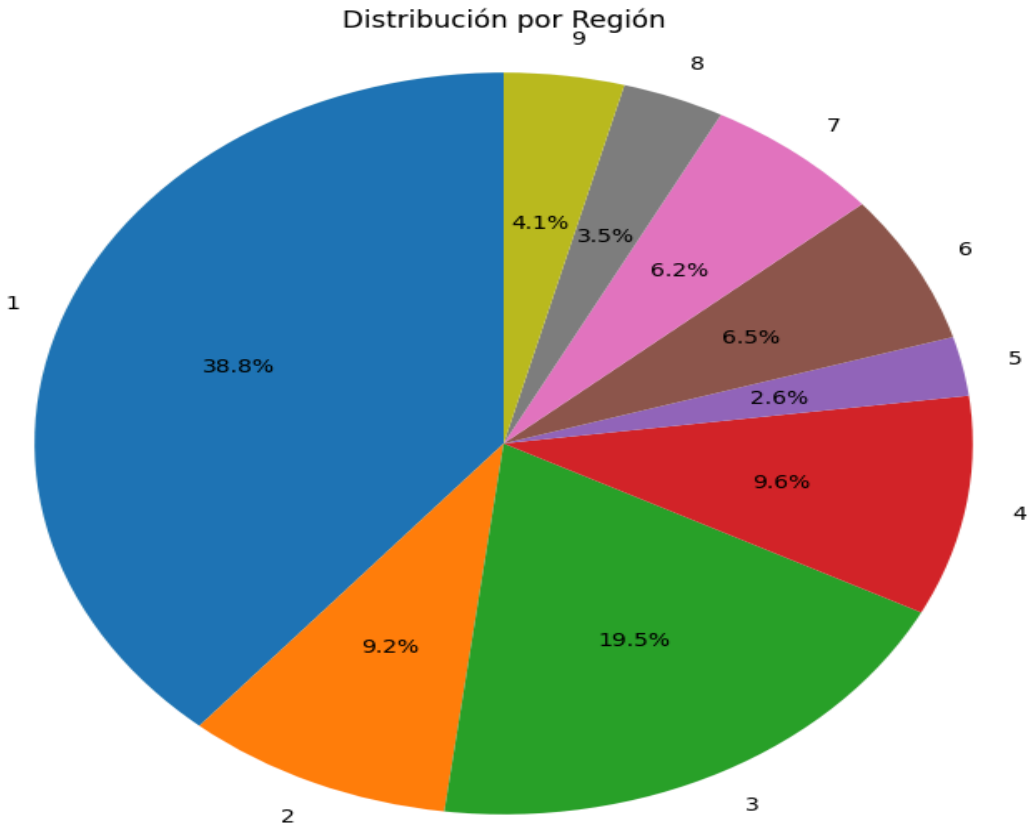
Un aspecto fundamental de esta variable es que el conjunto de datos está muy desbalanceado. Aproximadamente el 84.5% de las sesiones no resultaron en una compra (Revenue=False), mientras que solo el 15.5% sí lo hicieron (Revenue=True).Este desequilibrio debe ser considerado en cualquier modelo predictivo. Sin embargo, este desbalance debe ser interpretado con conocimiento del dominio. Una tasa de conversión del 15.5% por sesión puede parecer baja desde una perspectiva de balance de clases en ciencia de datos, pero en el contexto del comercio electrónico, es excepcionalmente alta.

Luego nos movimos por un conjunto de 4 variables categóricas que nos resultaron algo sospechosas por la poca posibilidad de interpretación que esta nos brinda.

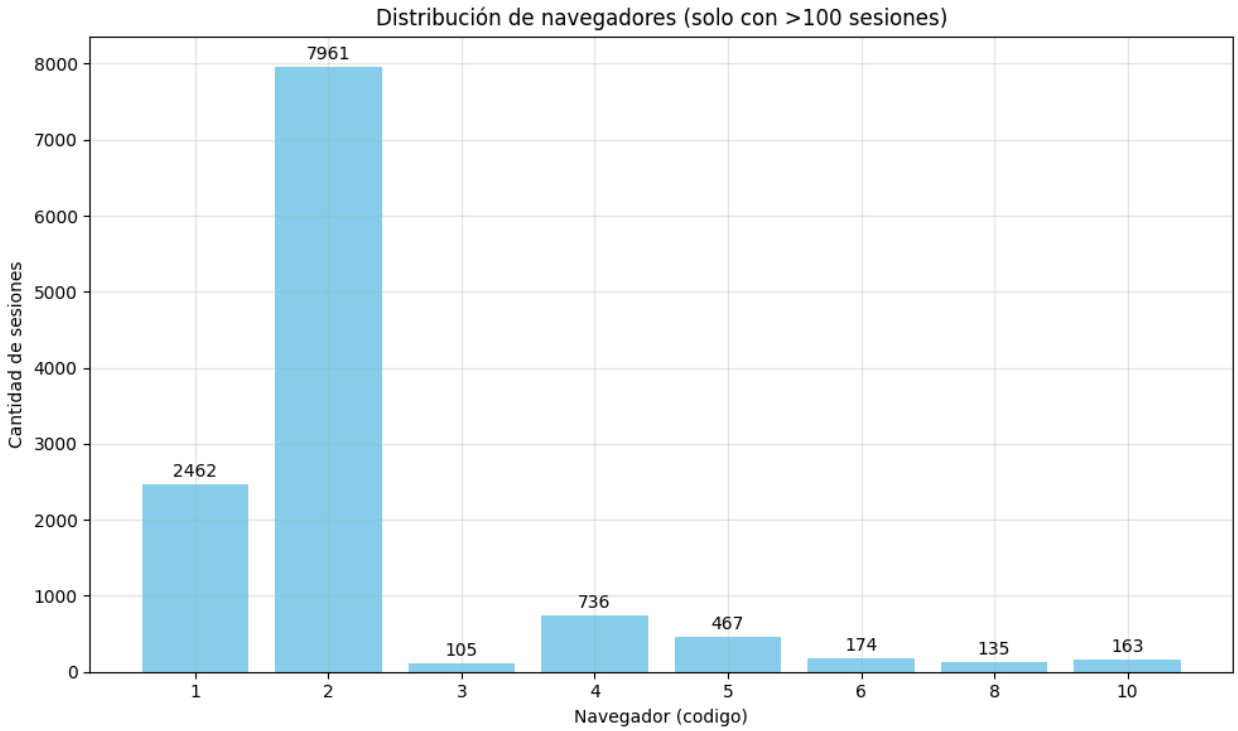


Notar que el 95 % de las muestras corresponden a los primeros 3 sistemas operativos.


Distribución de TrafficType (Tipo de Tráfico)
Cantidad de sesiones
Tipo de Tráfico (TrafficType)
0
500
1000
1500
2000
2500
3000
3500
4000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
2451
3913
2052
1069
260
444
40
343
42
450
247
1
738
13
38
3
1
10
17
198


Distribución por Región
1
2
3
4
5
6
7
8
9
38.8%
9.2%
19.5%
9.6%
2.6%
6.5%
6.2%
3.5%
4.1%

En la distribución de regiones podemos ver como tenemos una que se queda con el 38,8% y es la predominante,suponemos que al estar hecho el dataset por una Universidad de Turquía,estaría refiriéndose a esta región y las aledañas al continuar la distribución.



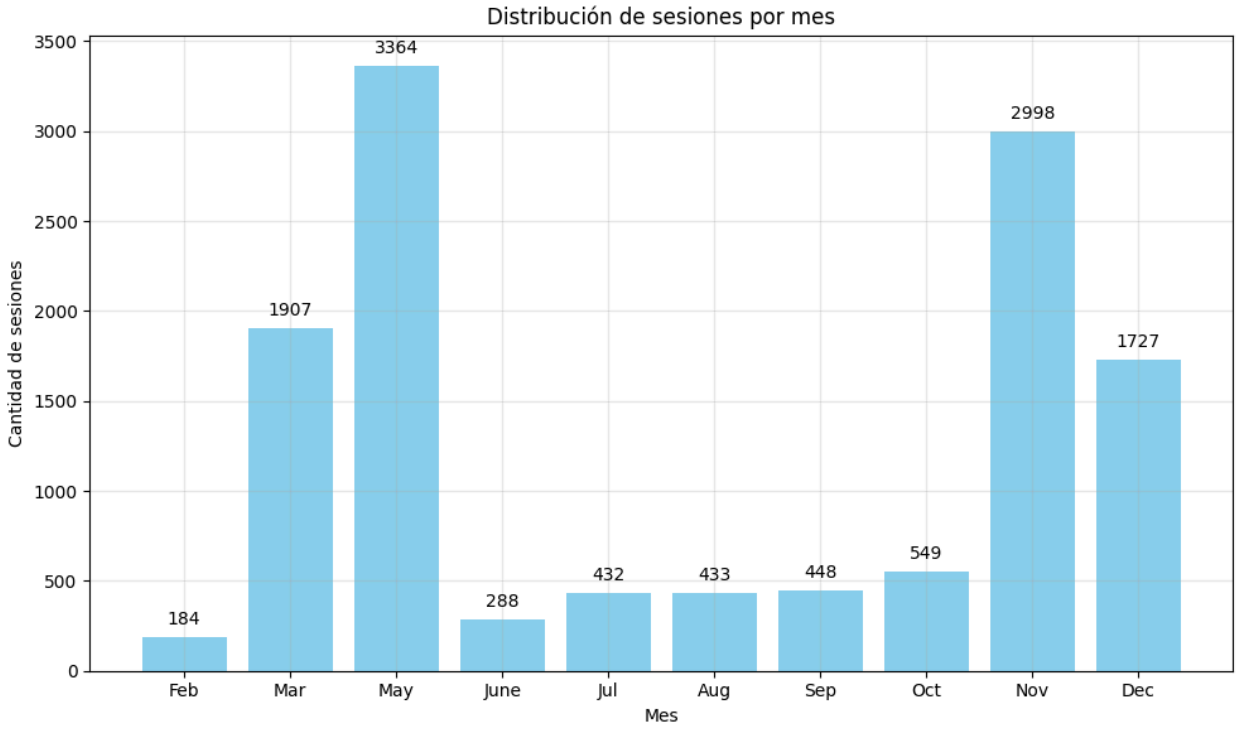
En este gráfico solo expusimos aquellos navegadores más significativos del dataset (aquellos con más de 100 sesiones registradas).

Hemos detectado una deficiencia crítica en la Calidad de Datos , ya que las variables categóricas nominales 'Browser', 'Región', 'Traffic Type' y 'Operative Systems' están codificadas numéricamente (ej. 1 a 8). Sin embargo, carecemos de los Metadatos o "diccionario" que aclaren qué representa cada valor (ej. “Browser 1 = Chrome”).Podríamos suponer a qué regiones se podría acercar pero no es confiable que nosotros supongamos los valores que se encuentran en este dataset. Esta ausencia de información, que afecta la dimensión de Validez de los datos, nos impide interpretar el significado semántico de estas características y, por lo tanto, limita severamente la profundidad del Análisis Exploratorio de Datos (EDA).

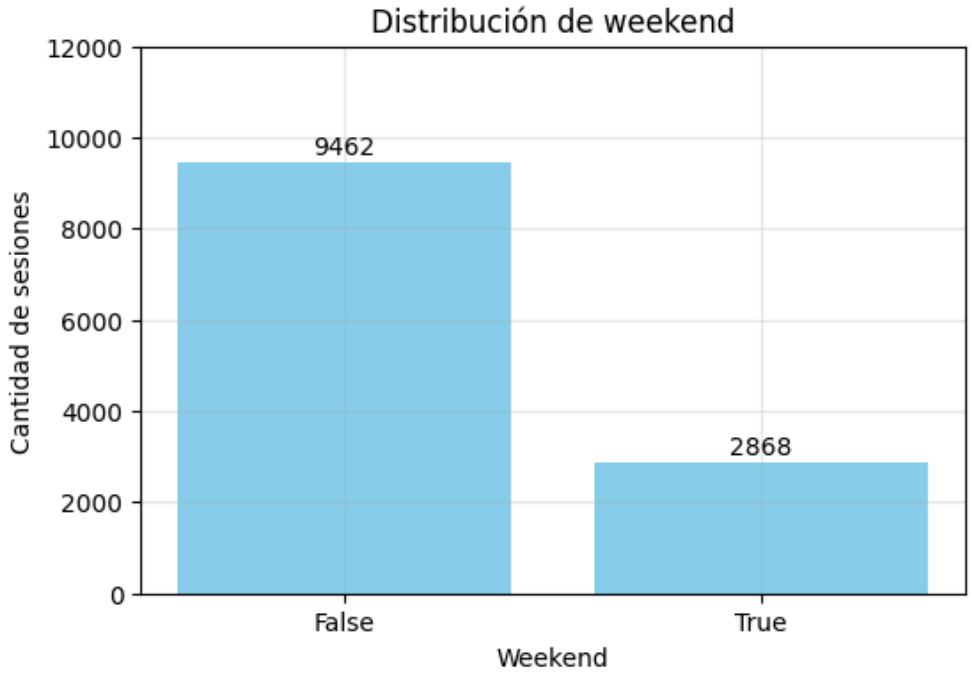
También hay otro tipo de variables categóricas que son los meses (“Month”) y si la sesión se realizó en un fin de semana(“Weekend”).

Además contamos con la variable día festivo (“Special Day”) que nos da una proximidad representada en un float.

Estos atributos proporcionan contexto sobre el momento y las circunstancias de la sesión.



Notar que en el dataset falta información sobre el mes Enero y Abril.

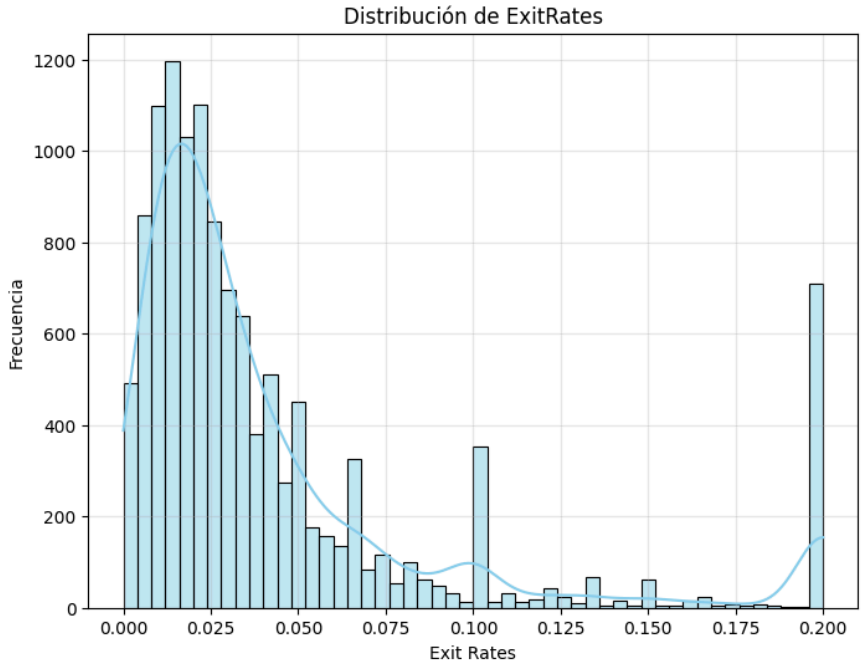


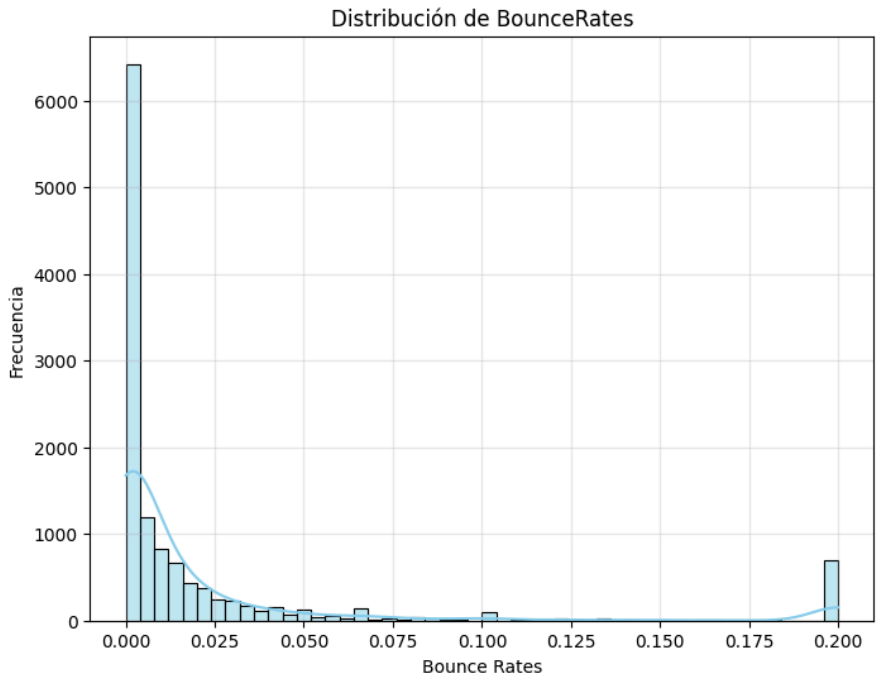
Special Day: Es una característica de ingeniería numérica que indica la proximidad de la fecha de la visita a un día especial de compras (por ejemplo, San Valentín, Día de la Madre). Es un valor de punto flotante entre 0 y 1, donde 0 indica ninguna proximidad y 1 indica la máxima proximidad. Su cálculo es sofisticado, ya que tiene en cuenta dinámicas del comercio electrónico como el tiempo entre la fecha del pedido y la entrega. Por ejemplo, para el día de San Valentín, el valor es distinto de cero entre el 2 y el 12 de febrero, y alcanza su máximo el 8 de febrero, no el 14, para dar tiempo al envío.

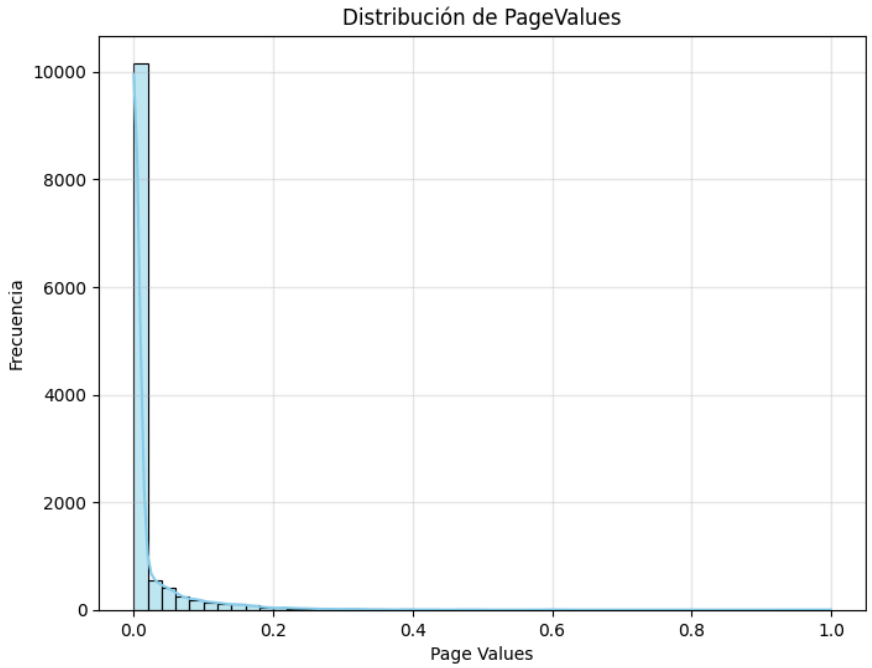
Month: El mes en que ocurrió la visita. Esta variable captura la estacionalidad (por ejemplo, compras navideñas en noviembre y diciembre). Es importante señalar que algunos análisis sugieren que faltan datos de dos meses en el conjunto de datos, lo que podría introducir un sesgo en los análisis basados en esta variable.

Weekend:Nos indica si la visita al sitio web se realizó durante el fin de semana,donde quizás el usuario cuenta con más tiempo de ocio.

Tenemos algunas variables que nos sirven de parámetro para ver cómo de eficiente es nuestra página con respecto a las visitas e interacciones con la misma.Nuestro dataset se apoya en las métricas de Google Analytics.







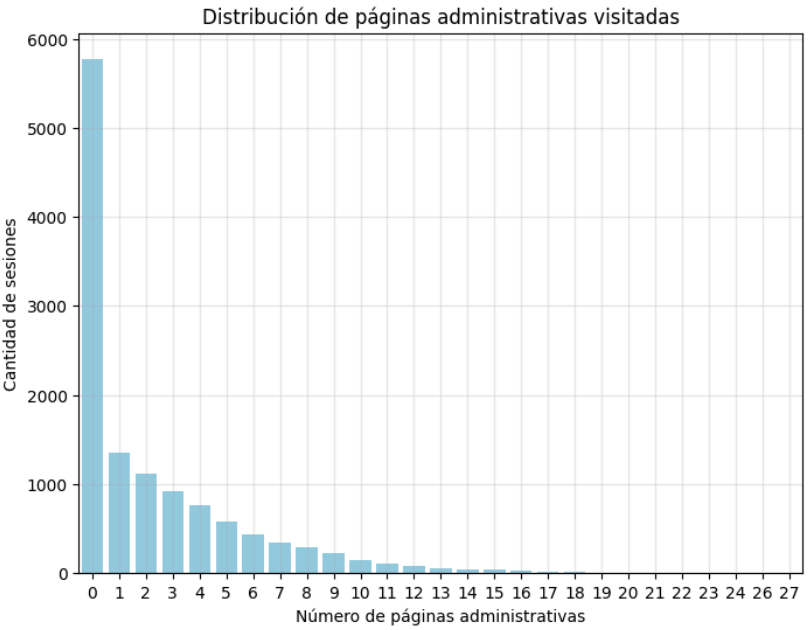
El análisis utiliza métricas clave de Google Analytics, como la Tasa de Rebote (Bounce Rates), que mide el desinterés del usuario (sesiones de una sola página), y la Tasa de Salida (Exit Rates), que identifica la última página de una visita. La métrica más

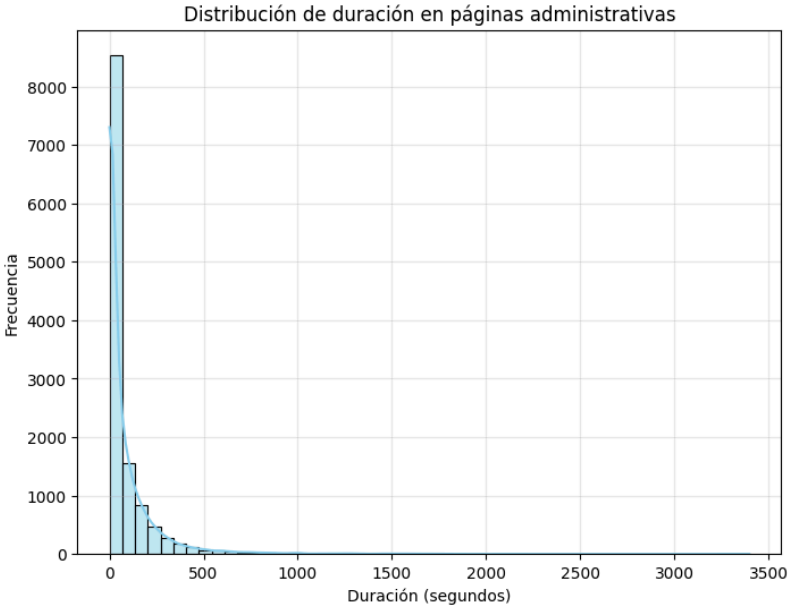
predictiva es el Valor de Página (Page Values), que asigna un valor económico a las páginas que contribuyeron a una conversión. Es fundamental entender que esta métrica es un *síntoma* y no una *causa* de los ingresos; dado que Page Values se calcula *basándose* en las transacciones, una sesión sin ingresos (Revenue=False) tendrá, por definición, un Page Values de 0. Por lo tanto, un Page Values alto no provoca la compra, sino que indica que el usuario ya está en un "camino dorado" con alta intención de conversión.

Tenemos además métricas que nos sirven para interpretar las interacciones realizadas por el usuario que se encuentra en la sesión.

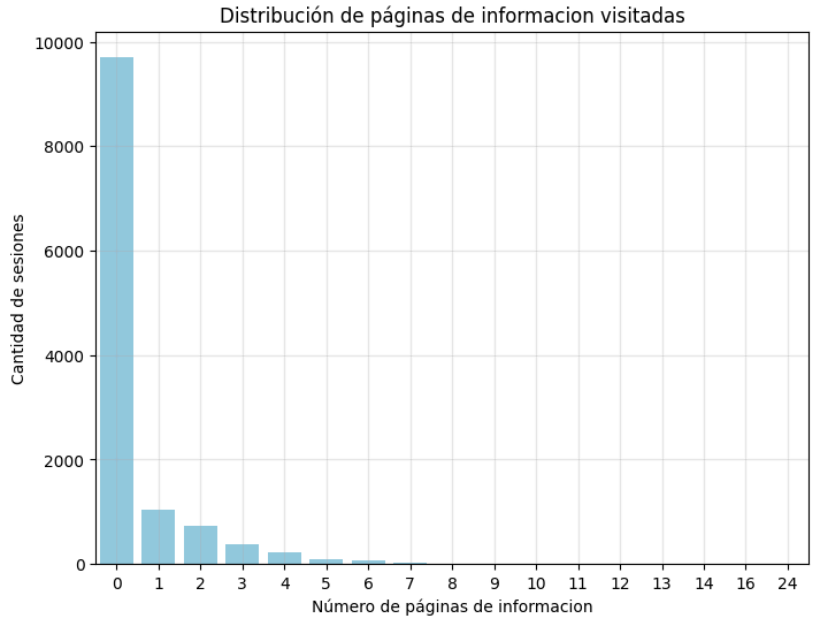
Las métricas de *clickstream* miden la interacción directa del usuario, agrupándolos en tres tipos de actividad.

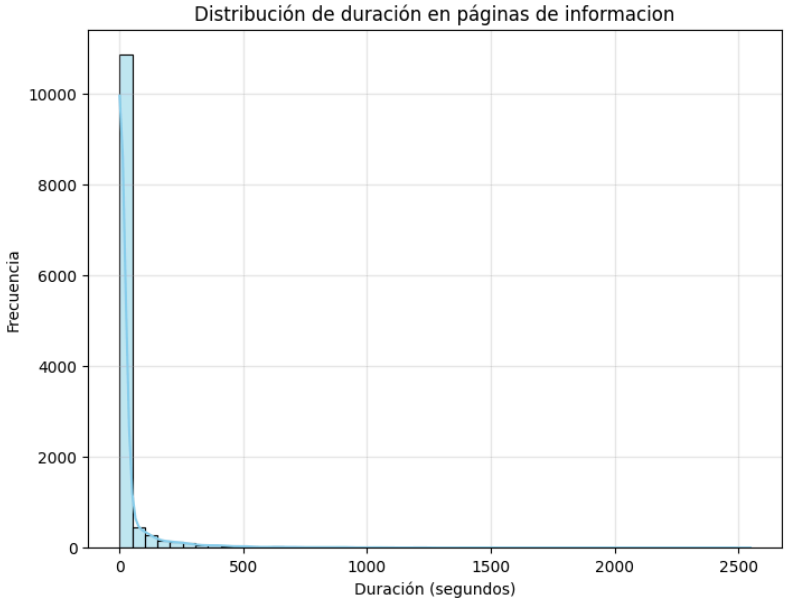
Administrative y Administrative Duration: Se refieren a las páginas relacionadas con la gestión de la cuenta, como el inicio de sesión, la visualización de pedidos anteriores o la configuración del perfil. Valores altos en estas métricas pueden indicar que un cliente existente está gestionando su cuenta, lo que podría ser un precursor de una nueva compra o una señal de un cliente leal.



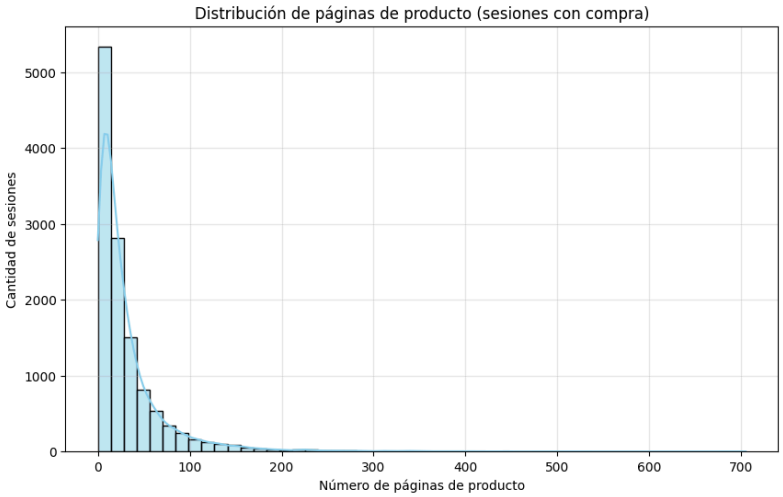


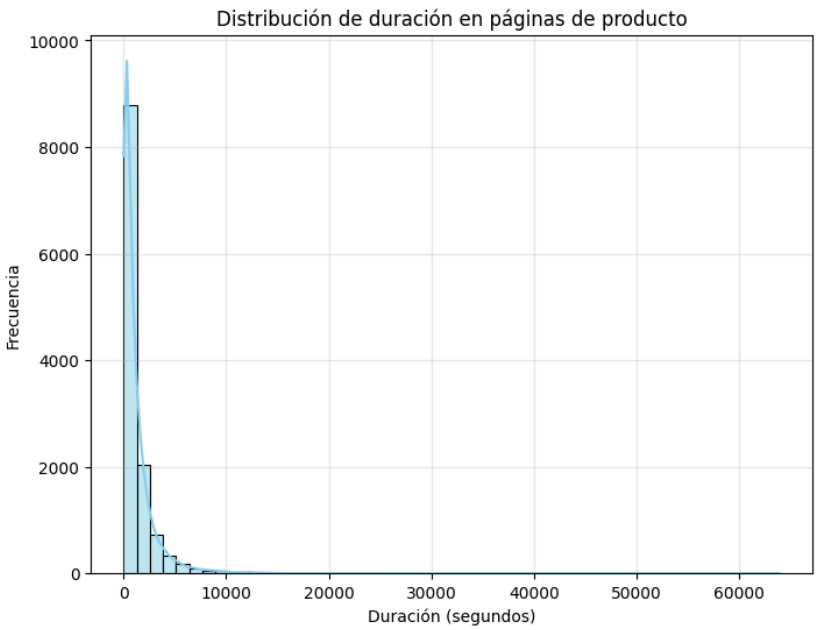
Informational y Informational Duration: Corresponden a páginas que proporcionan información general, como secciones de "Preguntas Frecuentes" (FAQ), políticas de envío o la página "Sobre nosotros". Una alta interacción con estas páginas sugiere que el usuario se encuentra en una fase de investigación o de construcción de confianza, evaluando la credibilidad y las condiciones del sitio antes de comprometerse a una compra.





Product Related y Product Related Duration: Representan el núcleo de la experiencia de compra. Incluyen páginas de categorías, páginas de detalles de productos y resultados de búsqueda. Se espera que esta sea la categoría más influyente en la intención de compra. La distinción entre el número de páginas (Product Related) y la duración (Products Related Duration) es crucial; un usuario que visita 50 páginas de productos en 60 segundos está probablemente "saltando" de un producto a otro sin mucho interés, mientras que un usuario que visita 10 páginas en 600 segundos está probablemente considerando cada producto con mayor detenimiento.





3. Calidad de Datos y Limpieza Inicial

Antes de realizar cualquier análisis, fue crucial evaluar la calidad de los datos. Este proceso reveló dos problemas principales: la existencia de registros duplicados y la presencia de serias inconsistencias lógicas.

3.1. Eliminación de Registros Duplicados

Se realizó una búsqueda de registros exactamente duplicados, un problema que afecta la dimensión de Exclusividad de los datos. Se identificaron 125 filas que eran copias idénticas de otras. Estos duplicados pueden distorsionar las estadísticas descriptivas y los conteos, por lo que se eliminaron, conservando solo la primera aparición de cada registro único. Esta acción redujo el dataset de 12,330 a 12,205 observaciones, asegurando que cada fila representa una sesión distinta.

3.2. Detección de Inconsistencias Lógicas

El segundo paso fue verificar la coherencia interna, un aspecto clave de la Calidad de Datos. Se detectó una inconsistencia lógica recurrente: existían registros donde un usuario supuestamente había visitado una o más páginas de un tipo (ej. ProductRelated > 0), pero el tiempo total registrado en esa categoría era exactamente cero (ProductRelated_Duration = 0).

Este problema violaba la Validez de los datos, ya que no es lógico visitar una página sin que transcurra un tiempo mínimo. Se encontraron 1,078 registros (un 8.74% del dataset) con este error, repartidos entre las categorías Administrative, Informational y ProductRelated. Estos valores cero se trataron como datos problemáticos, equivalentes a datos faltantes (missing data).

4. Imputación de Inconsistencias Lógicas

Simplemente eliminar estos 1,078 registros (casi el 9% del dataset) habría resultado en una pérdida significativa de información y podría haber introducido sesgos en la muestra.

Para preservar la información de estos registros,se descartó la eliminación de filas y se optó por la imputación de datos para corregir los valores de 0.0 duración.Utilizamos la media ajustada heurísticamente,qué consistió primero en obtener una duración promedio apoyándonos en los registros consistentes y en segundo lugar, imputar los valores inconsistentes con el 50% de dicha duración promedio.El 50% es un balance entre no ignorar que se visitan páginas y reconocer que fueron sesiones atípicas;este recurso nos da a entender que estas sesiones fueron anómalas o cortas.

Esta decisión nos pareció apropiada ya que toma un balance conservador,ya que así evitamos la sobreestimación y el sesgo que la media simple habría generado.

Al finalizar este proceso, el conjunto de datos quedó libre de duplicados y con sus inconsistencias lógicas corregidas, resultando en un dataset íntegro y coherente, listo para la siguiente fase del análisis.

5. Transformación de Características (Encoding)

El último paso de la preparación fue realizar transformaciones de características. Como vimos en la materia, la mayoría de los algoritmos de machine learning operan con datos numéricos, por lo que fue preciso convertir las variables categóricas y booleanas a un formato numérico adecuado.

5.1. Conversión de Variables Binarias

En primer lugar, se realizó una conversión simple de las variables binarias. Las columnas Weekend y Revenue, que estaban almacenadas como valores booleanos, se transformaron a un formato numérico binario: 1 para "True" y 0 para "False". Este paso

las estandariza para el análisis.

5.2. Codificación de Variables Categóricas.

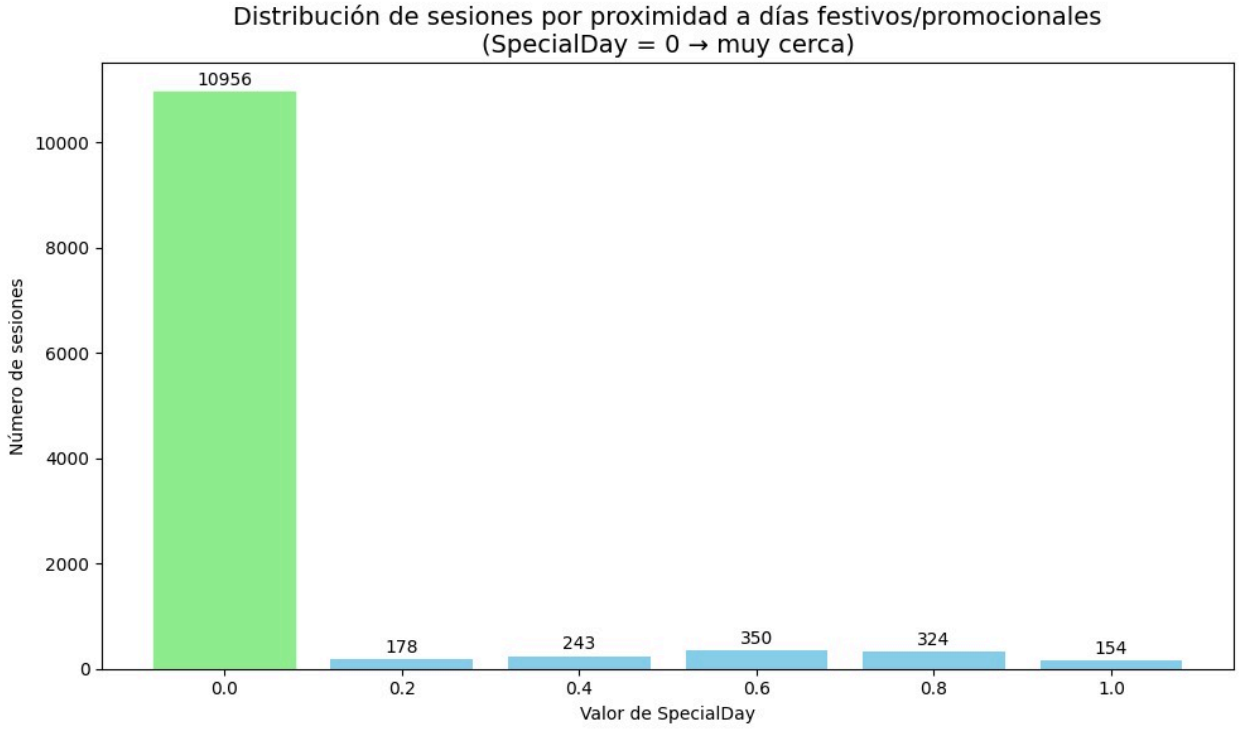
Además se analizó la variable categórica VisitorType.Se evaluó realizar un One-Hot Encoding.Sin embargo, una de las desventajas conocidas es que aumenta la dimensionalidad del conjunto de datos. En pos de mantener la máxima simplicidad del modelo y evitar este incremento en el número de columnas, se tomó la decisión de excluir la variable VisitorType del análisis final en la transformaciones.

Decidimos transformar las variables 'Browser', 'Traffic Type', 'Región' y 'Operative Systems' a categóricas porque, aunque computacionalmente estén registradas como números , su naturaleza estadística es cualitativa nominal. Como vimos en clase, si las tratáramos como numéricas, los algoritmos de Machine Learning podrían "inducir que hay un orden que no es tal", asumiendo erróneamente una relación de magnitud (ej. Región 4 > Región 1) que no existe.

HIPÓTESIS Y RESOLUCIONES

Hipótesis 1:

"La gran mayoría de las sesiones ocurren en fechas cercanas a días festivos, lo que indica que el tráfico está fuertemente impulsado por las oportunidades de mercado relacionadas a estas fechas."

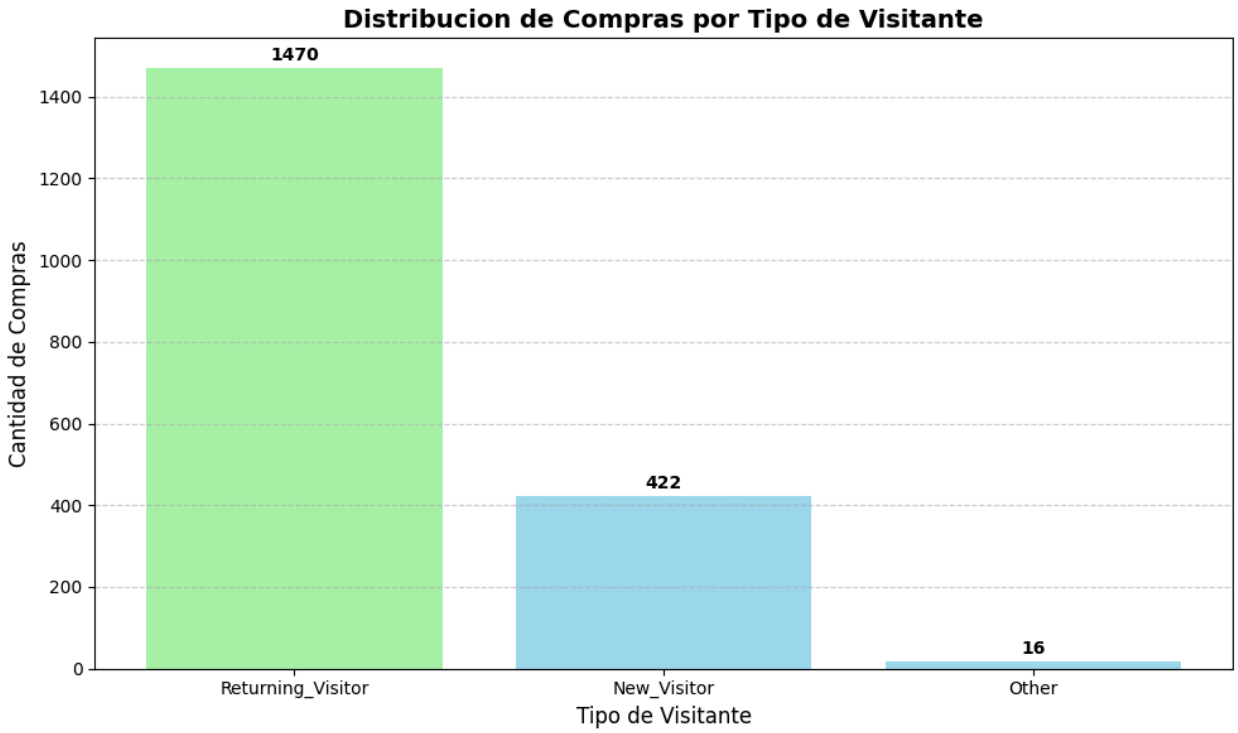


El análisis arrojó un resultado contundente: la distribución de `Special Day` está extremadamente desbalanceada. De un total de 12,205 sesiones, 10,956 (es decir, el 89.8%) se agrupan bajo el valor 0.0, que indica la máxima proximidad a un evento especial. Los demás valores (0.2, 0.4, 0.6, 0.8 y 1.0) son comparativamente residuales y representan, en conjunto, solo al 10.2% de los datos.

Esta abrumadora concentración validó la suposición inicial, demostrando que el sitio web es muy atractivo cuando ocurren eventos relacionados a días festivos y las sesiones se disparan considerablemente.

Hipótesis 2:

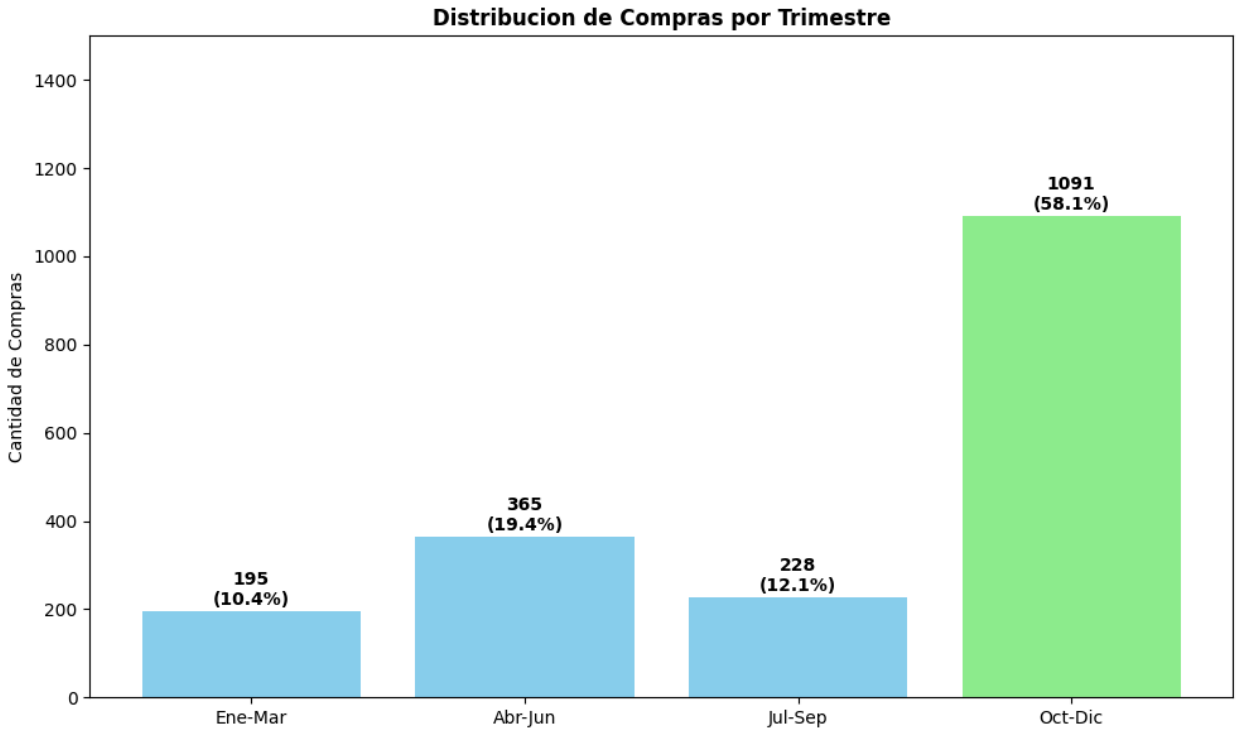
"La mayoría de las compras están asociadas a sesiones de usuarios que ya visitaron el sitio y tienen experiencia previa."



El análisis de las compras por tipo de visitante muestra un claro dominio por parte de los 'Returning_Visitor', quienes concentran la gran mayoría de las transacciones (77.0%). Esto sugiere que la retención de clientes es un factor crucial para el negocio. Los 'New_Visitor' representan un segmento secundario importante (22.1%), aunque su contribución es significativamente menor, mientras que la categoría 'Other' (0.8%) es prácticamente insignificante en el volumen total de compras.

Hipótesis 3:

"El último trimestre del año es donde se registran más sesiones finalizadas en compra."



Para esta hipótesis, elegimos el test de Chi-cuadrado (no-paramétrico) porque es la herramienta correcta para comparar conteos en diferentes categorías, que en nuestro caso son los cuatro trimestres. Queríamos ver si la cantidad de compras que observamos en cada trimestre se desviaba de una distribución uniforme (es decir, si todos los trimestres tuvieran la misma cantidad de compras).

El test de chi-cuadrado pone en juego las siguientes hipótesis:

H0 = Los 4 trimestres tienen igual cantidad de compras.

H1 = Existen diferencias significativas entre los trimestres.

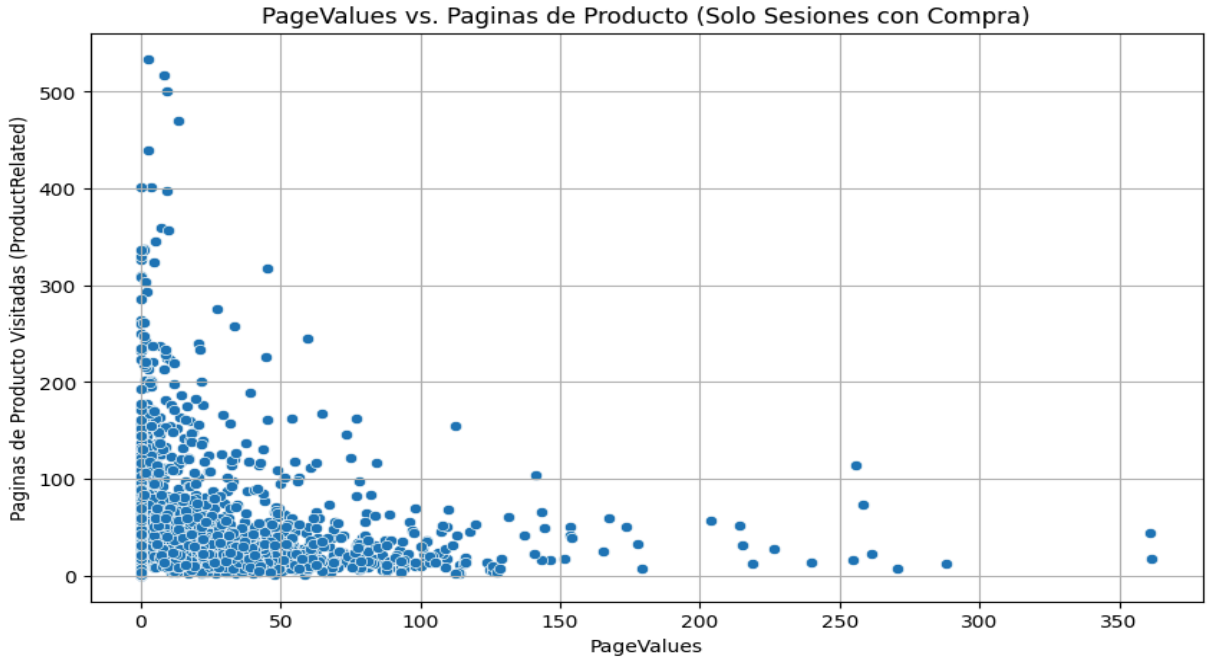
El test verifica si son realmente uniformes la cantidad de compras o si al menos un trimestre es significativamente diferente a los demás, pero no sabe cual. Pero al haber observado el gráfico de la distribución de compras por trimestre y teniendo la frecuencia en término de porcentaje del 58.1% para el último trimestre, concluimos que el mayor número de compras en el último trimestre es una diferencia real y estadísticamente significativa.

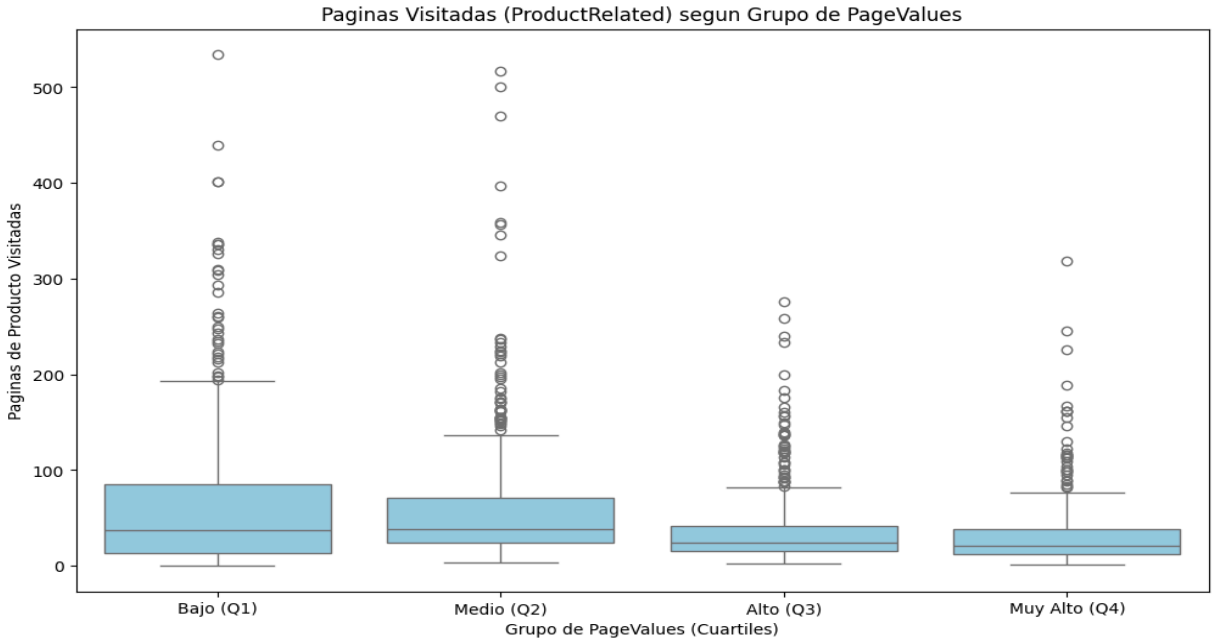
El test de Chi-cuadrado de bondad de ajuste arrojó un p-valor = 0.000000 muy inferior al nivel de significancia de alfa = 0.05, lo que nos lleva a rechazar la hipótesis nula de que las compras se distribuyen uniformemente a lo largo del año.

Esto confirma que existen diferencias estadísticamente significativas entre los trimestres, y al observar la distribución (con el 70.3% de las compras en "Oct-Dic"), validamos la hipótesis de investigación: el último trimestre es, de hecho, el período donde se registra la mayor y más significativa cantidad de sesiones finalizadas en compra.

Hipótesis 4:

"Entre las sesiones que resultaron en compra, y un alto valor de PageValues se corresponde con que se visitan muchas páginas o hubo una exploración profunda del sitio"





La hipótesis NO se cumple.

¿Por qué?

Los datos muestran que las sesiones con mayor valor de PageValues (las más valiosas) NO son las que visitan más páginas de producto.

En el gráfico de dispersión, los puntos con PageValues altos (hacia la derecha) están en su mayoría en la parte baja del eje vertical (pocas páginas visitadas).

El boxplot confirma que, en promedio, los grupos con PageValues "Alto" y "Muy Alto" tienen menos páginas visitadas que los grupos "Bajo" y "Medio".

En otras palabras:

No es necesario visitar muchas páginas para generar un alto valor.

Muchas sesiones valiosas (alta PageValues) ocurren después de una navegación rápida o focalizada, no tras una exploración larga.

“No necesitamos que los usuarios naveguen mucho para que se cumplan los objetivos trazados(Goals). De hecho, las compras más valiosas a menudo vienen de quienes encuentran rápido lo que buscan. Las sesiones registradas refieren a una experiencia rápida y directa, no en hacer que la gente ‘exploré’ más.”

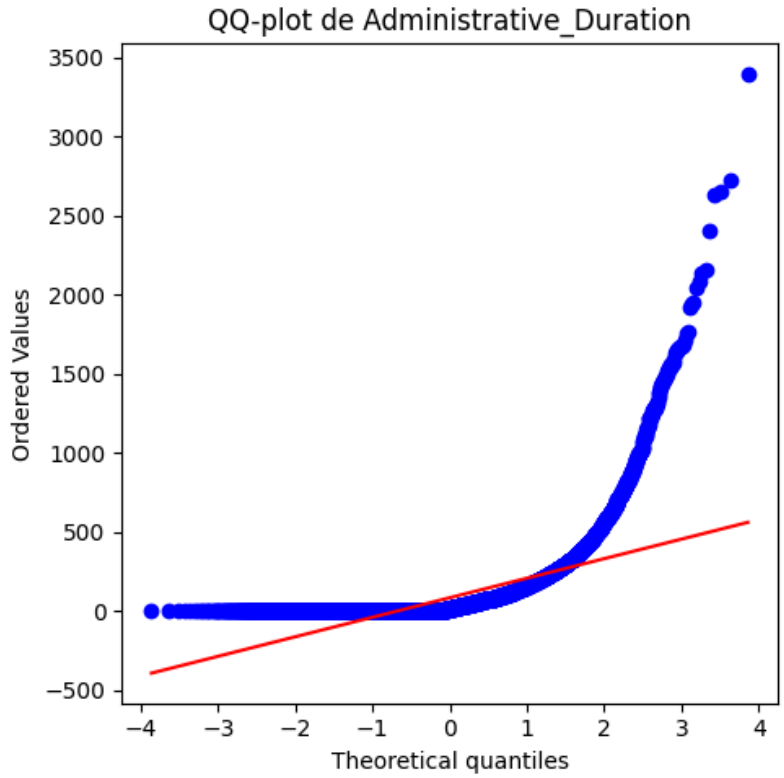
Hipótesis 5:

"El tiempo promedio de las sesiones en páginas administrativas (inicios de sesión, configuración,proceso de pago, etc) es el mismo para sesiones que terminan en compra y sesiones que no resultan en compra."

Se sospecha que si una sesión pasa un tiempo excesivo en secciones Administrativas y además la sesión no termina en compra, quizá porque el usuario está "atascado" (fricción operativa).Por lo tanto, se buscó comparar la distribución de la variable cuantitativa Administrative_Duration entre dos grupos independientes: sesiones con Revenue=True (Compra) y sesiones con Revenue=False (No Compra).

Para comparar una variable cuantitativa entre dos grupos independientes, el test paramétrico estándar es el Test t de Student.Sin embargo, los tests paramétricos tienen supuestos que deben cumplirse. El más importante para el Test t es el supuesto de normalidad. El análisis exploratorio y los tests formales revelaron que la variable

Administrative_Duration no sigue una distribución normal; de hecho, presenta un fuerte sesgo, como lo demuestran las diferencias entre las medias y las medianas.Al no cumplirse el supuesto de normalidad, se descartó el Test t.



En su lugar, se seleccionó correctamente la alternativa no paramétrica: el Test de Mann-Whitney U. Este test compara las medianas (o más precisamente, las distribuciones) de dos muestras independientes sin asumir una distribución específica.Las hipótesis planteadas fueron:

H0 = El tiempo promedio en páginas administrativas es igual en ambos grupos.

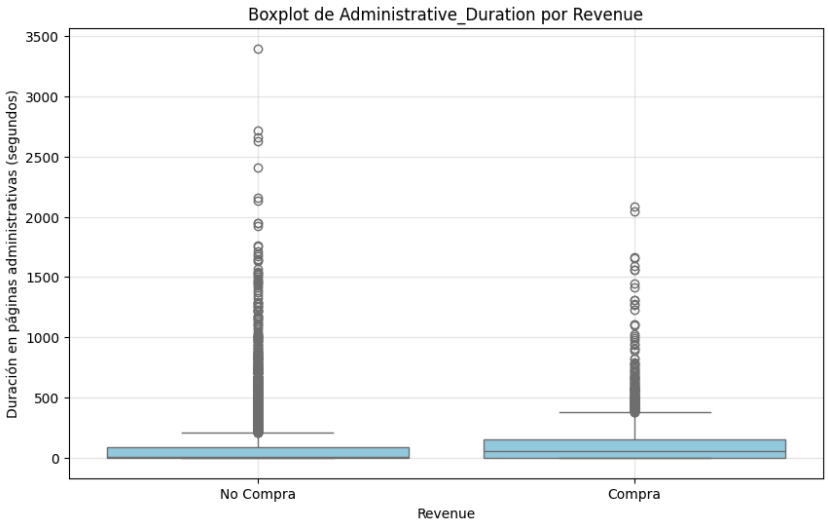
H1 = El tiempo promedio en páginas administrativas es diferente entre los grupos.

El Test de Mann-Whitney U arrojó un resultado contundente.El rechazo de la hipótesis nula nos permite afirmar que existe una diferencia estadísticamente significativa en el tiempo que las sesiones pasan en páginas administrativas entre aquellas que terminan en compra y aquellas que no.Para entender esta diferencia nos apoyamos en estadísticas descriptivas como puede ser la mediana(más robusta).

Mediana (Sesiones que SÍ Compran): 56.80 segundos

Mediana (Sesiones que NO Compran): 3.00 segundos

El tiempo en secciones administrativas no representa un obstáculo, sino un paso necesario y característico del proceso de compra exitoso. Los usuarios que compran deben, por definición, interactuar con estas páginas (iniciar sesión, registrarse, ingresar datos de pago, confirmar la orden, etc.), acumulando así una duración significativamente mayor. Por el contrario, la gran mayoría de sesiones que no compran (mediana de 3 segundos) apenas interactúan con estas secciones.



Hipótesis 6 :

"Existen perfiles de compradores definidos o los compradores se comportan de manera similar."

H0 = No existen grupos diferenciados. Todas las sesiones de compradores provienen de una única distribución de comportamiento.

H1 = Existen 3 grupos (clusters) que se corresponden con perfiles de comprador teóricos: "Perfil Explorador", "Perfil Metódico o Desconfiado" y "Perfil Directo".

Perfil 1: El Explorador.

Este es un cliente que, antes de finalizar su compra, invirtió una cantidad significativa de tiempo navegando activamente por el catálogo de productos (un valor alto en

'ProductRelated_Duration'). A diferencia de un comprador impulsivo, este perfil compara y se informa sobre los productos en detalle. Su sesión, aunque culminó en una venta, se caracterizó por esta fase de exploración exhaustiva, mientras que el tiempo dedicado a la gestión de su cuenta (Administrativo) o a leer páginas de ayuda (Informativo) fue mínimo.

Perfil 2: El Comprador Metódico o desconfiado.

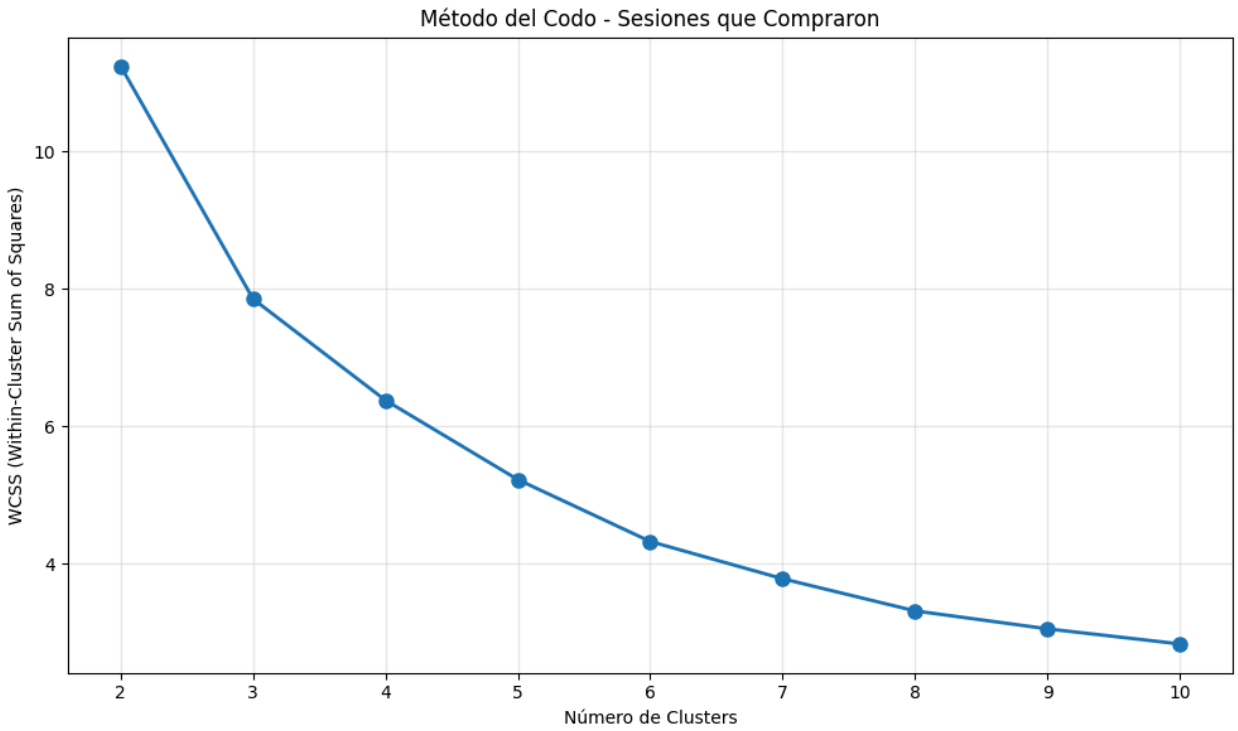
Este perfil describe a un cliente que, aunque terminó comprando, necesitó validar la confianza o resolver dudas antes de hacerlo. Su tiempo de sesión se distribuyó principalmente entre páginas informativas (leyendo 'Preguntas Frecuentes', 'Políticas de Devolución', etc.) y páginas administrativas (quizás creando una cuenta o verificando su perfil). Lo notable es que completaron la compra *sin* dedicar mucho tiempo a navegar por el catálogo de productos. Esto sugiere que ya sabían qué querían, pero su prioridad era asegurarse de que el sitio fuera confiable o entender las condiciones de la transacción antes de confirmar el pago.

Perfil 3: El Comprador Directo o eficiente

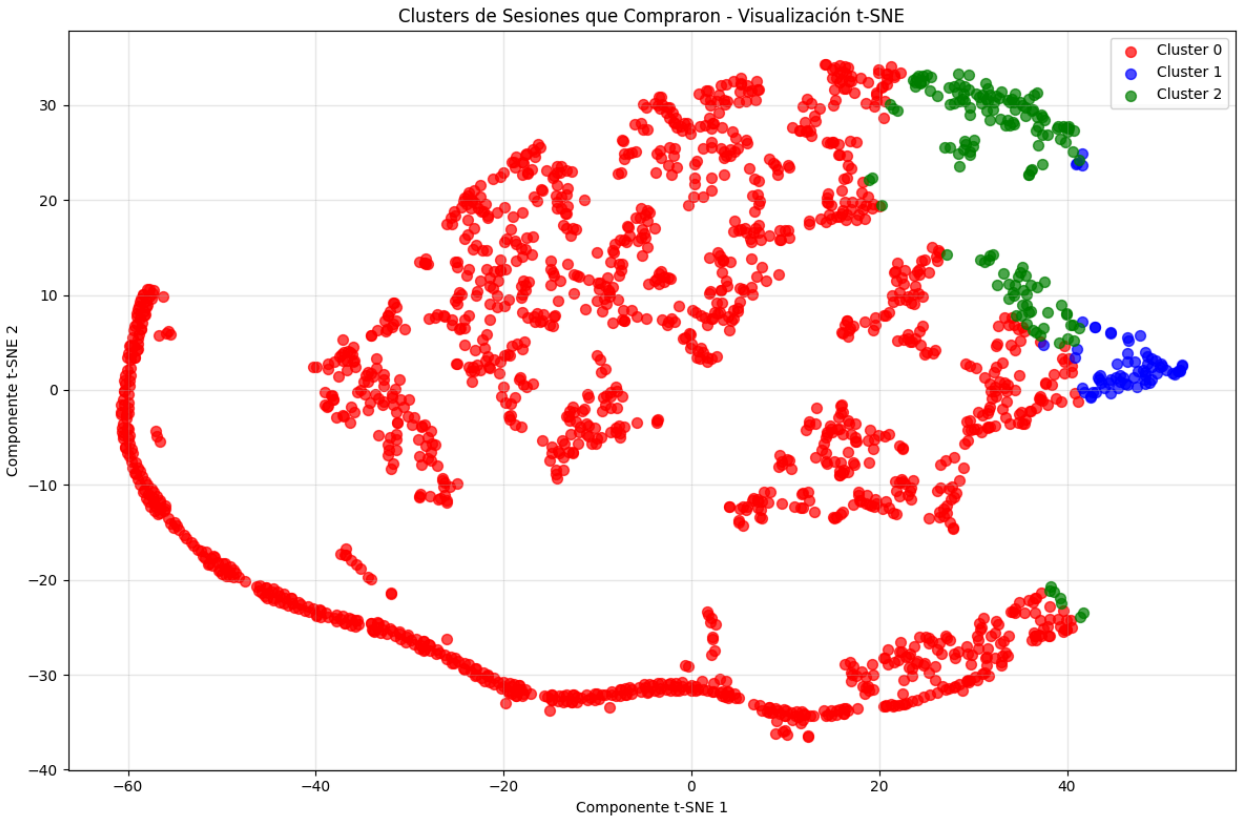
Este es el perfil del cliente más eficiente. Su sesión se centró casi exclusivamente en las páginas administrativas (un valor alto en 'Administrative_Duration'), lo que en el contexto de una compra exitosa significa que fueron directos al proceso de pago (checkout). Este cliente no necesitó explorar productos ni leer páginas de ayuda. Es la representación de una compra decidida: el cliente entró, inició sesión, pagó y salió. Esto es típico de un cliente recurrente que realiza una compra de reposición o que ya tenía su decisión tomada de antemano.

El algoritmo seleccionado, K-Means, es un método basado en distancia. Como las variables Administrative_Duration, Informational_Duration y ProductRelated Duration están en la misma escala (segundos), el escalado MinMaxScaler asegura que todas las variables tengan el mismo peso en el cálculo de la distancia, evitando que una domine a las otras.

Elección del Número de Clusters (K):Aunque la hipótesis sugería K=3, se utilizó el Método del Codo (Elbow Method) para una validación empírica. Este método gráfica la inercia para diferentes valores de K.El gráfico mostró una clara inflexión ("codo") en K=3.Esta evidencia empírica, combinada con la fuerte interpretabilidad teórica, solidifica la elección de K=3 como la óptima.



En la interpretación de los Perfiles (Centroides). El algoritmo K-Means identificó 3 grupos, y el análisis de sus centroides (características promedio) se alinea perfectamente con los perfiles teorizados:



Cluster 0 (1681 sesiones):Perfil Explorador -> ProductRelated_Duration: 0.025 (Más Alto)Administrative_Duration: 0.021.Informational_Duration: 0.010. Este es el grupo principal. Pasan la mayor parte de su tiempo en páginas de productos, coincidiendo con el comportamiento de "navegación y exploración" del perfil.

Cluster 1 (68 sesiones): Perfil Metódico o Desconfiado -> Informational_Duration: 0.313 (Drásticamente más Alto)Administrative_Duration: 0.081.ProductRelated_Duration: 0.073. Un grupo pequeño pero distintivo que pasa un tiempo abrumador en páginas informativas (FAQ, "Acerca de", políticas de envío). Esto se alinea perfectamente con usuarios nuevos o cautelosos que buscan información y confianza.

Cluster 2 (159 sesiones): Perfil Directo -> Administrative_Duration: 0.175 (Notablemente más Alto)ProductRelated_Duration: 0.056.Informational_Duration: 0.037.Este grupo pasa la mayor parte de su tiempo en tareas administrativas. Esto sugiere usuarios que van "directo al grano": inician sesión y proceden directamente al checkout.

Para probar que estas agrupaciones son estadísticamente significativas y no solo un artefacto del algoritmo, se realizó el tests paramétrico : ANOVA (Análisis de Varianza)

H0: Las medias de duración (para cada feature) son iguales en los tres clusters.H1: Al menos una media es diferente.Resultados: Para las tres variables (Administrative, Informational y ProductRelated), el pvalor fue 0.000.Se rechaza la H0 en todos los casos. Existe una diferencia estadísticamente significativa entre los clusters.

Pero como mencionamos anteriormente, ANOVA es un test paramétrico. Por lo cual se requiere que se cumplan sus supuestos para ser verídico en términos de probabilidad y resultados. Se comprobó el supuesto de Homocedasticidad a través del test de Levene para las tres variables de interés. Y resultó que no se cumplía la igualdad de varianzas entre ellas, por lo tanto se descartaron todas las conclusiones sacadas con ANOVA. Luego, se hizo uso de su alternativa no-paramétrica : test de Kruskal-Wallis.

Este nuevo test arrojó resultados iguales para las 3 variables, se concluye que existen diferencias significativas entre las medianas en al menos un cluster. Pero nuestro trabajo ahora es identificar y decir cual cluster es el realmente distinto. Por ello se utilizaron múltiples test U para hacer comparaciones por pares e identificar los clusters.

La Hipótesis 6 se valida contundentemente. El análisis no solo identificó 3 grupos distintos, sino que demostró que estos grupos se alinean perfectamente con los perfiles de comprador teorizados. El Método del Codo justificó empíricamente la elección de K=3. Las diferencias de comportamiento (tiempo en páginas) entre los perfiles "Explorador", "Metódico/Desconfiado" y "Directo" son estadísticamente significativas y no producto del azar.

CONCLUSIÓN

La exploración del dataset “Online Shoppers Intention” nos permitió entender el comportamiento del usuario en un entorno de e-commerce, con el objetivo final de predecir la intención de compra en tiempo real, imitando la intuición de un vendedor experimentado en una tienda física.

Como se planteó en la introducción, el desafío central del e-commerce es convertir visitas en compras, especialmente cuando los usuarios parecen interesados pero abandonan el carrito. Nuestro análisis confirmó que el comportamiento digital del usuario (su “lenguaje corporal digital”) contiene señales claras que pueden ser utilizadas para intervenir en el momento preciso y aumentar la conversión.

Implicancias de Negocio

- Personalización en tiempo real: Conociendo el perfil del usuario (de acuerdo con su comportamiento actual), se puede intervenir con ofertas, mensajes o guías personalizadas.
- Optimización de campañas: Invertir en marketing durante los últimos 3 meses del año, pero también crear eventos en otros momentos para suavizar la estacionalidad.
- Fidelización: Diseñar programas de lealtad y experiencias personalizadas para convertir a los nuevos visitantes en recurrentes.
- Experiencia de usuario: Aunque el proceso de pago no es un problema, la exploración previa sí lo es. Se debe mejorar la navegación, la búsqueda y la presentación de productos para reducir la fricción en las etapas iniciales.

Este análisis demuestra que el comportamiento del usuario en línea no es aleatorio, sino predecible. Al entender los patrones de navegación, los tipos de visitantes y los momentos críticos, las empresas pueden transformar el “abandono de carritos” en oportunidades de venta.

REFERENCIAS

1. https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset.
2. https://www.kaggle.com/datasets/imakash3011/online-shoppers-purchasing-intention-dataset .
3. Informe basado en Deep Research para investigación a fondo del dominio.