

ESTRATEGIA DE CREENCIAS

Para "Horizon Digital"

1. Introducción

"Horizon Digital", una empresa de e-commerce especializada en productos electrónicos, se encuentra en un punto crucial de planificación estratégica. La dirección busca fundamentar su estrategia para el próximo año fiscal no en la intuición, sino en un análisis de datos robusto.

Como Científico de Datos de la compañía, he recibido el encargo de realizar un **análisis de 360 grados** de nuestros datos de clientes y ventas. El objetivo principal es transformar el conjunto de datos ecommerce_data.csv en *insights* accionables y modelos predictivos que guíen las decisiones en áreas clave como finanzas, marketing y operaciones.

Este notebook documentará todo el proceso, desde la exploración inicial de los datos hasta la generación de pronósticos y recomendaciones estratégicas.

2. Objetivos del proyecto

El análisis se estructurará en cinco "misiones" principales, cada una diseñada para responder a una pregunta de negocio específica:

★ **Misión 1: Entendiendo el Negocio (Análisis Exploratorio de Datos - EDA)**

- **Objetivo:** Obtener una visión general de la situación actual.
- **Pregunta clave:** ¿Quiénes son nuestros clientes (demografía) y cómo se comportan (interacción y gasto)?

★ **Misión 2: Prediciendo el Valor del Cliente (Regresión)**

- **Objetivo:** Estimar el gasto total histórico (TotalSpending) de un cliente.
- **Pregunta clave:** ¿Qué factores (ingresos, tiempo en el sitio, etc.) influyen más en el gasto de un cliente y podemos predecir su valor para optimizar los costes de adquisición?

★ Misión 3: Identificando Clientes Potenciales (Clasificación)

- **Objetivo:** Predecir la probabilidad de que un cliente se suscriba al boletín (IsSubscribed).
- **Pregunta clave:** ¿Qué perfil de cliente es más propenso a suscribirse, permitiendo al equipo de marketing enfocar sus campañas de captación?

★ Misión 4: Creando Campañas Personalizadas (Clustering)

- **Objetivo:** Segmentar la base de clientes en grupos homogéneos.
- **Pregunta clave:** ¿Podemos identificar "personas" (ej. "VIP", "Potenciales", "Leales") basadas en sus ingresos y gastos para diseñar estrategias de marketing personalizadas?

★ Misión 5: Pronosticando las Ventas Futuras (Series Temporales)

- **Objetivo:** Prever el volumen de ventas mensuales para el próximo año.
- **Pregunta clave:** ¿Cuál es la demanda esperada para los próximos 12 meses, con el fin de optimizar la gestión de inventario y la logística?

3. Carga y limpieza de datos

Calidad General de los Datos

La primera gran noticia es que el dataset está en excelente forma en cuanto a integridad. Con 2.500 registros:

- **No tenemos valores nulos (NaN)** en ninguna de las 10 columnas.
- Esto nos ahorra un paso significativo de limpieza e imputación de datos. Podemos proceder directamente al análisis y la preparación de características.

Acciones de Preparación Requeridas

Hemos identificado dos tareas de pre-procesamiento que son indispensables antes de continuar con las misiones de modelado:

- **Conversión de Fechas (Crítico para Misión 5):** La columna LastPurchaseDate está actualmente como tipo object (texto). Para poder realizar el análisis de series temporales (Misión 5), es fundamental que la convirtamos a un formato datetime.
- **Codificación de Categóricas (Crítico para Misiones 2 y 3):** La columna Gender también es de tipo object. Para utilizarla en los modelos de regresión y clasificación, necesitaremos transformarla en variables numéricas (probablemente usando *One-Hot Encoding*).

Primeros Insights del Negocio (Análisis Descriptivo)

El método .describe() ya nos da pistas muy valiosas sobre nuestros clientes:

Perfil del Cliente: Tenemos una base de clientes muy diversa. La edad (Age) va de 18 a 74 años (media de 45.7) y los ingresos anuales (AnnualIncome) van de 20k a 150k (media de 84.15k). Esta diversidad es ideal para la segmentación (Misión 4).

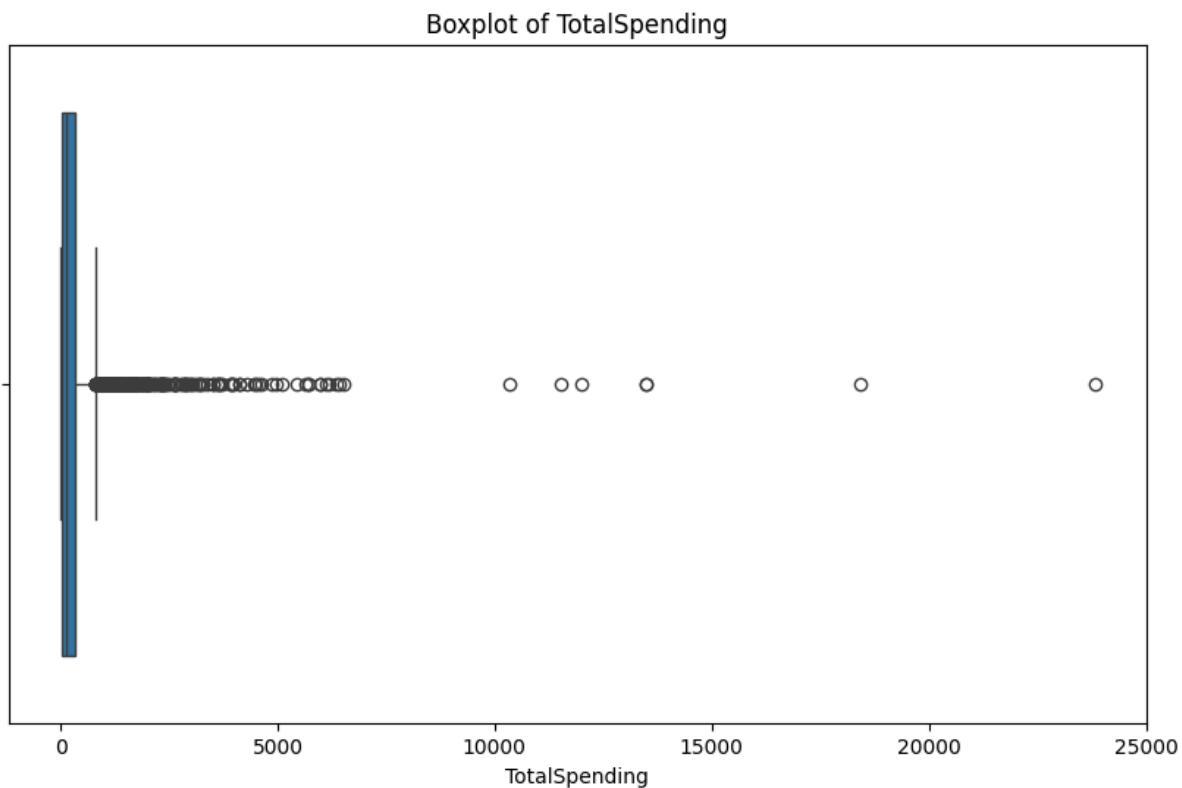
El Hallazgo Clave (Asimetría de Gastos): Existe una enorme disparidad en el gasto de los clientes.

- **TotalSpending:** La media es 400, pero el 75% de los clientes (percentil 75) ha gastado solo 347 o menos. Sin embargo, ¡el gasto máximo (max) es de 23,819!
- **LastPurchaseAmount:** Vemos lo mismo. La media es 53, pero el máximo es 4,032. Esto nos quiere decir que la mayoría de nuestros clientes gasta cantidades moderadas, pero tenemos un pequeño grupo de clientes "VIP" (o *outliers*) que gastan muchísimo más. Deberemos tener esto en cuenta para el modelo de regresión (Misión 2), ya que estos valores extremos pueden sesgar el resultado. Quizás necesitemos escalar o transformar esta variable.
- **Suscripciones (Para Misión 3):** La media de IsSubscribed es 0.4136. Esto significa que el 41.36% de nuestros clientes están suscritos. Es una proporción bastante equilibrada (cercana al 50/50), lo cual es excelente para entrenar un modelo de clasificación sin demasiado desbalanceo de clases.
- **Comportamiento en el Sitio:** El tiempo promedio en el sitio (TimeOnSite) es de 11.4 minutos y los clientes dejan, en promedio, 4.5 artículos en el carrito (ItemsInCart).

En resumen, los datos están limpios, pero requieren transformación de tipos. Analíticamente, la gran asimetría en el gasto es el *insight* más importante hasta ahora y será un punto central en nuestras misiones de regresión y clustering.

3.1 Gestión de nulos, duplicados y outliers

Como he dicho, hemos tenido suerte al no tener nulos y tampoco hay duplicados. Al comprobar los outliers confirma y magnifica lo que sospechábamos al ver el `.describe()`. Estas son las conclusiones clave que sacamos:



La Gran Desigualdad en el Gasto:

El hallazgo más evidente es que la distribución del TotalSpending está **extremadamente sesgada hacia la derecha** (sesgo positivo).

- **La "Caja" Comprimida:** La caja azul, que representa al 50% central de nuestros clientes (del percentil 25 al 75), está completamente aplastada a la izquierda, muy cerca del cero. Esto nos dice que la gran

mayoría de nuestros clientes (al menos el 75%) gasta cantidades relativamente pequeñas y muy similares entre sí.

- **Los "Outliers" (Puntos Clave):** Los puntos individuales (los círculos) que se extienden hacia la derecha no deben ser tratados como errores. **Estos son nuestros clientes más importantes.**

Los "Outliers" son nuestros Clientes "VIP"

Estos puntos no son ruido; son la señal más fuerte que tenemos. Representan a un pequeño subconjunto de clientes "VIP" cuyo gasto es órdenes de magnitud superior al del cliente promedio. Vemos clientes con gastos que superan los 10,000, 15,000 e incluso se acercan a los 24,000.

Implicaciones para Nuestras Misiones

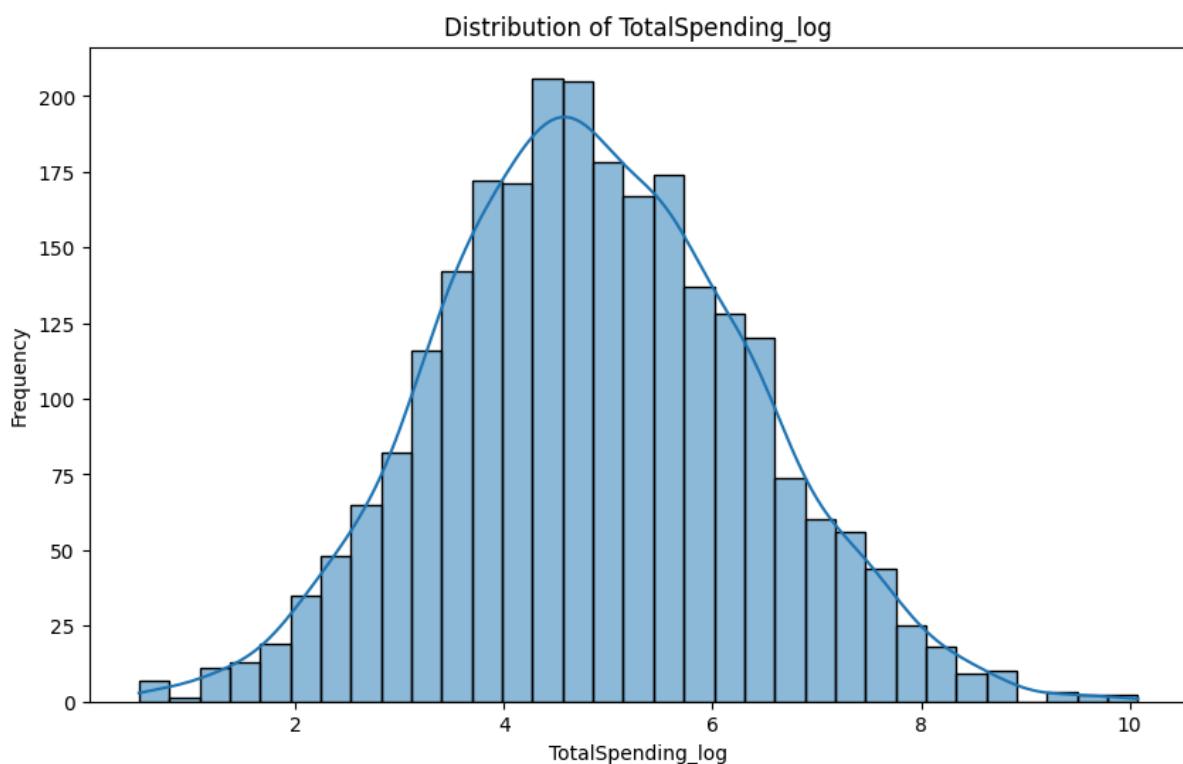
Esta visualización tiene consecuencias directas para nuestros siguientes pasos:

- **Para la Misión 2 (Regresión):** Entrenar un modelo para predecir TotalSpending será un desafío. Los modelos de regresión, incluido RandomForestRegressor, pueden verse fuertemente influenciados por estos valores extremos. Es muy probable que el modelo se esfuerce por predecir estos valores altos y, como resultado, sea menos preciso para la gran mayoría de clientes "normales".
- **Acción Requerida:** Tendremos que **considerar seriamente transformar nuestra variable objetivo TotalSpending** (por ejemplo, aplicando una transformación logarítmica, $\log(x+1)$) antes de entrenar el modelo. Esto ayudará a "comprimir" la escala y hacer la distribución más simétrica, mejorando el rendimiento del modelo.
- **Para la Misión 4 (Clustering):** Esta estructura de datos es perfecta para la segmentación. Es casi seguro que el algoritmo K-Means identificará a estos "outliers" como un clúster separado y de alto valor. Esto valida la necesidad de la dirección de crear campañas personalizadas; el marketing para un cliente de 20,000 no puede ser el mismo que para uno de 200.

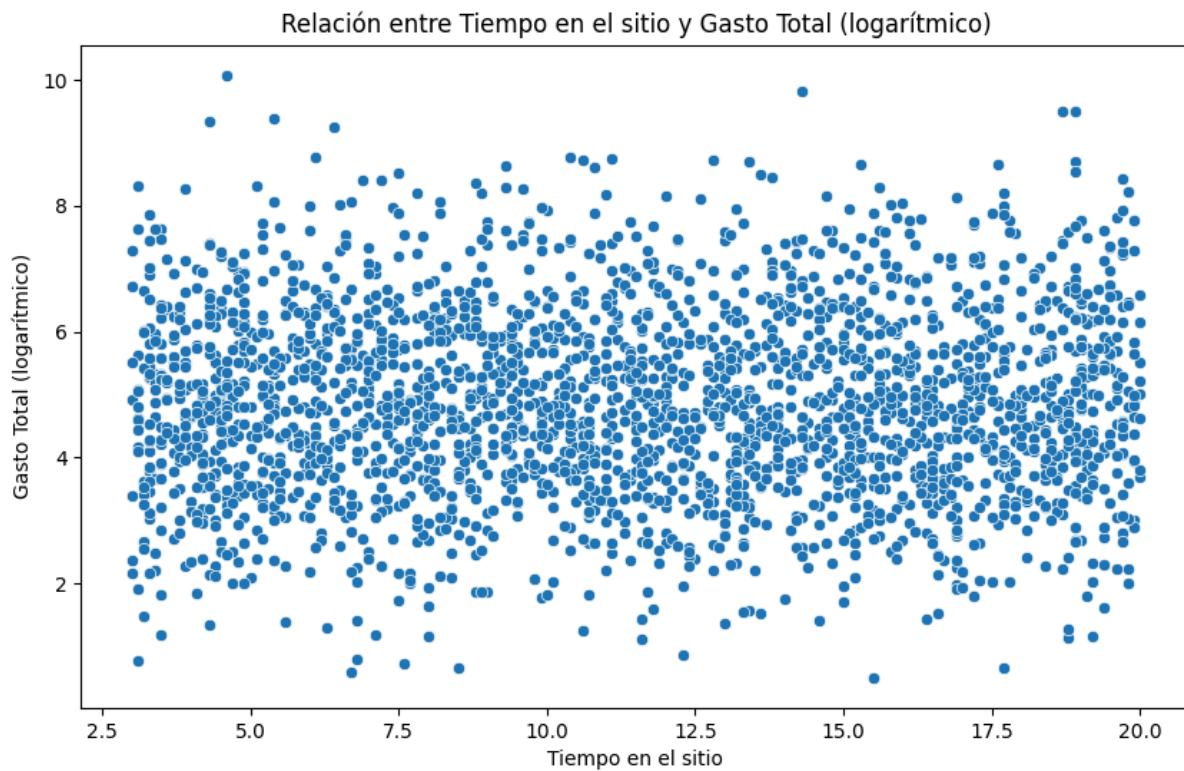
4. Transformación Logarítmica

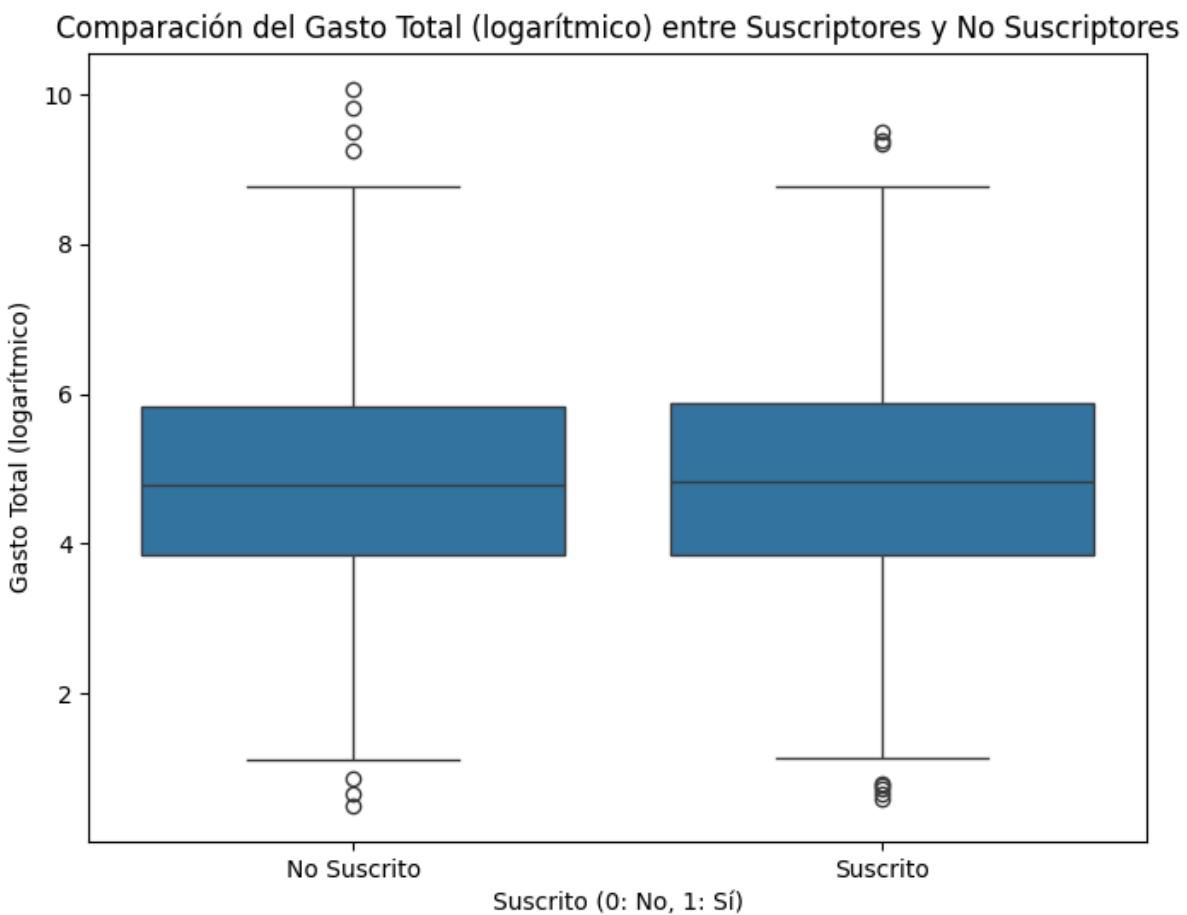
¿Por qué es la mejor técnica?

1. **"Comprime" la Escala:** Esta transformación "acerca" a los clientes VIP al resto, sin borrar la información de que son los que más gastan. Reduce drásticamente el sesgo y hace que la distribución se parezca mucho más a una campana de Gauss (distribución normal).
2. **Estabiliza los Modelos:**
 - **Para la Misión 2 (Regresión):** Los modelos de regresión funcionan mucho mejor cuando la variable objetivo (TotalSpending) tiene una distribución normal. Al entrenar el modelo para predecir $\log(\text{TotalSpending})$, el error (RMSE) será más bajo y las predicciones más fiables.
 - **Para la Misión 4 (Clustering):** K-Means se basa en la distancia. Sin esta transformación, los VIP crearían su propio clúster solo por estar "lejos", y la segmentación del resto de clientes (el 99%) sería muy pobre. Al transformar la variable, permitimos que el algoritmo encuentre patrones reales (ej. "altos ingresos pero bajo gasto") en lugar de ser cegado por los valores absolutos.



¡Éxito de la Transformación! Hemos resuelto con éxito el problema de los outliers y el sesgo extremo. La transformación logarítmica (`np.log1p`) ha funcionado de maravilla. Hemos convertido una distribución totalmente asimétrica (vista en el boxplot) en una distribución normal (la clásica campana). Los datos ahora están centrados simétricamente alrededor de un valor medio (aproximadamente 4.5-5.0 en la escala logarítmica).





Conclusión 1: El Tiempo en el Sitio NO influye en el Gasto Total

- **Gráfico:** Relación entre Tiempo en el sitio y Gasto Total (Scatter plot).
- **Observación:** La respuesta a la pregunta "¿Pasan más tiempo los clientes que más gastan?" es un **no claro**.
- **Análisis:** El gráfico de dispersión es una "nube de puntos" sin forma, densa y sin una tendencia discernible. No existe una correlación visible (ni positiva ni negativa) entre TimeOnSite y TotalSpending_log. Un cliente que pasa 5 minutos en el sitio puede tener un gasto total tan alto o bajo como un cliente que pasa 20 minutos.
- **Insight de Negocio:** Esto es un hallazgo clave. Sugiere que la estrategia de la empresa no debe centrarse únicamente en "aumentar el engagement medido en tiempo", sino en la calidad de esa interacción. El tiempo por sí solo no se traduce en ventas.

Conclusión 2: Los Suscriptores NO gastan más que los No Suscriptores

- **Gráfico:** Comparación del Gasto Total (logarítmico) entre Suscriptores y No Suscriptores (Boxplot).

- **Observación:** La respuesta a la pregunta "¿Los suscriptores gastan más?" también parece ser **no**.
- **Análisis:** Los dos diagramas de caja son casi idénticos. La mediana (la línea central), el rango intercuartílico (la "caja") y la distribución general del gasto total logarítmico son virtualmente los mismos para el grupo "Suscripto" (1) y el "No Suscripto" (0).
- **Insight de Negocio:** Este es quizás el insight más sorprendente hasta ahora. Contrario a lo que se podría esperar, estar suscripto al boletín no está correlacionado con un mayor gasto total. Esto le dice al equipo de marketing que, si bien el boletín puede servir para otros propósitos (lealtad, comunicación), **actualmente no es un diferenciador de los clientes de alto valor**. Esto hace que la Misión 3 (predecir quién se suscribe) sea interesante, pero también nos dice que debemos buscar otros factores (como los que veremos en la regresión) para predecir el gasto.

A continuación lo que hemos hecho es lo que hemos hecho es **partir nuestros 2.500 clientes en dos grupos**:

1. **Set de Entrenamiento (80% - 2000 clientes):** Son los "apuntes" que usará el modelo para "estudiar" y aprender los patrones (X_{train} , y_{train}).
2. **Set de Prueba (20% - 500 clientes):** Es el "examen final" que el modelo nunca ha visto. Lo usaremos para comprobar si realmente ha aprendido algo (X_{test} , y_{test}).

La única conclusión es que **ya estamos listos para una evaluación justa**. Esto nos asegura que el modelo no solo "memoriza" los datos, sino que aprende a predecir el gasto de clientes que nunca ha visto.

Luego a la hora de entrenar el modelo y empeza con las predicciones hemos obtenido este resultado:

Root Mean Squared Error (RMSE) : 0.6379

Mean Absolute Error (MAE) : 0.5006

R-squared (R^2) : 0.8113

Lo que significa que significa que estas métricas indican que tu modelo RandomForestRegressor tiene un rendimiento aceptable para predecir el gasto total logarítmico, con un R-cuadrado sólido que muestra que gran

parte de la variabilidad en el gasto está siendo explicada por las variables predictoras.

5. Modelo Random Forest

El modelo solo usa UNA característica: Nuestro modelo RandomForestRegressor ha decidido que, para predecir el Gasto Total (TotalSpending_log), solo necesita mirar una cosa. Todos los demás datos son, en su opinión, casi irrelevantes.

Esa única característica es LastPurchaseAmount (Monto de la Última Compra), que **explica el 87% de la predicción del modelo**.

Las características que creímos importantes (Ingresos, Edad, Tiempo en el Sitio) son prácticamente ignoradas.

El "Por qué" (El problema de la circularidad): El modelo ha aprendido la siguiente lógica: "Si quieras saber cuánto ha gastado un cliente en total, la mejor pista es ver cuánto gastó la última vez".

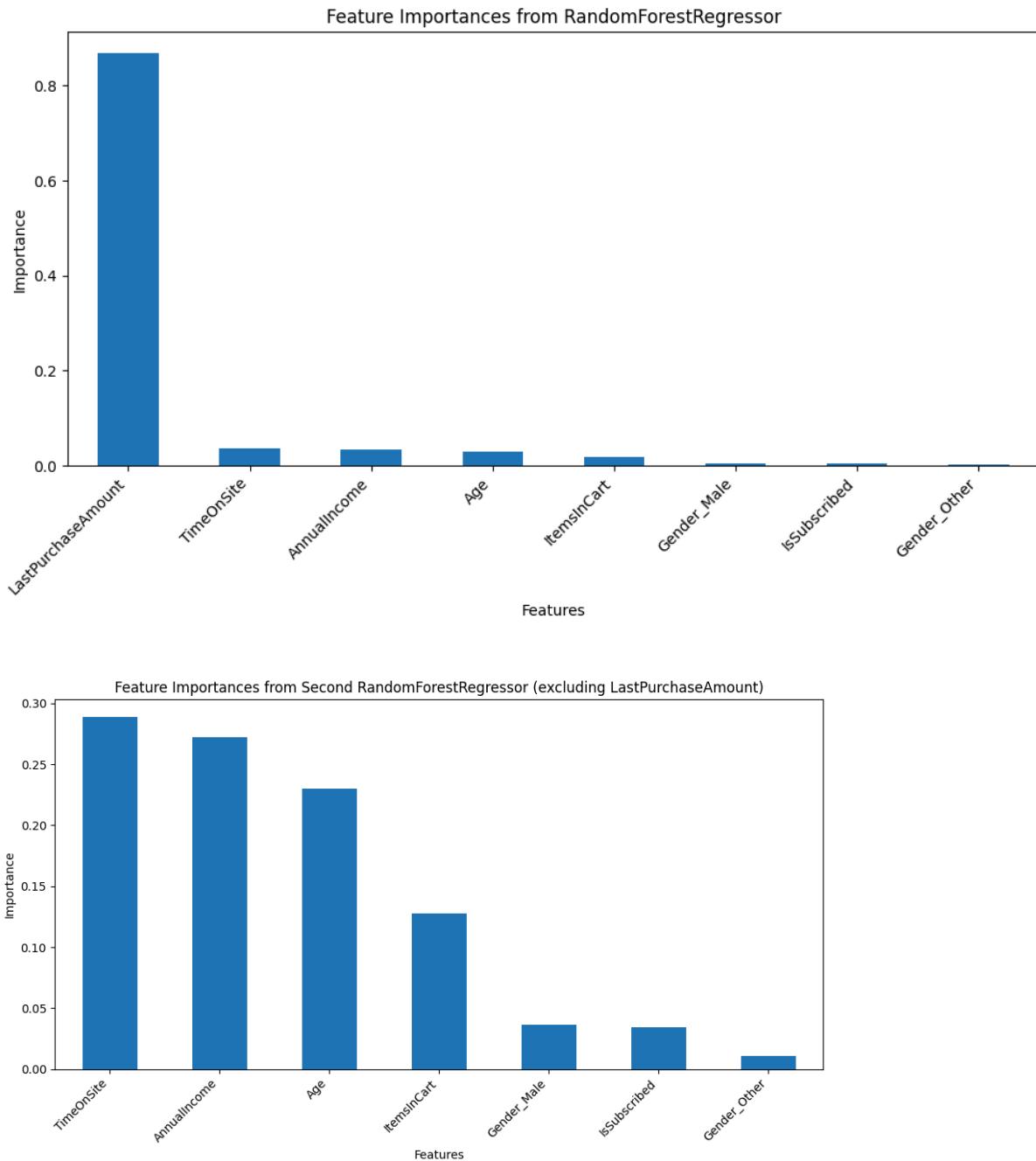
Esto tiene sentido estadísticamente (clientes que gastan mucho, gastan mucho), pero es un **razonamiento circular** y no nos ayuda mucho con el objetivo de negocio.

Conclusión para la Dirección de Finanzas: Este es el punto clave:

- **Lo que el modelo HACE:** Predice el gasto total basándose en el gasto reciente.
- **Lo que el modelo NO HACE:** No nos dice qué tipo de cliente (edad, ingresos, género) es el que gasta más.

Básicamente, le estamos diciendo al equipo de finanzas: "Un cliente gastará mucho si... gasta mucho". Esto no les da ninguna palanca estratégica. No pueden "aumentar los ingresos anuales" de un cliente, pero sí pueden *dirigir campañas* a clientes con un perfil específico.

En resumen: El modelo es probablemente preciso (lo veremos con el RMSE), pero no es *útil* para entender los *impulsores* del gasto. La variable LastPurchaseAmount es tan potente que "oculta" el efecto del resto de características.



Hasta ahora hemos entrenado nuestro primer modelo de Random Forest Regressor para predecir el gasto total transformado (TotalSpending_log) y obtuvimos métricas de evaluación (RMSE, MAE, R-cuadrado) que indicaron un **buen rendimiento predictivo general**, explicando una parte significativa de la varianza en el gasto logarítmico.

Al analizar la importancia de las características en este primer modelo, **hemos descubierto que el monto de la última compra (LastPurchaseAmount) es, con una diferencia abrumadora, el factor más importante** para predecir el gasto total. Esto sugiere que el historial de compras recientes es un predictor muy potente del gasto acumulado.

Para entender la influencia de otros factores, **hemos entrenado un segundo modelo de regresión excluyendo el monto de la última compra**. Las métricas de este segundo modelo mostraron un **rendimiento predictivo mucho más bajo** (incluso con un R-cuadrado negativo), confirmando la gran importancia de LastPurchaseAmount.

Finalmente, al examinar la importancia de las características en el segundo modelo, **hemos identificado que, entre las variables restantes, el tiempo en el sitio (TimeOnSite), los ingresos anuales (AnnualIncome) y la edad (Age) son los factores con mayor influencia relativa** en el gasto total. Esto nos da insights sobre otros aspectos del cliente (comportamiento en el sitio y demografía) que, aunque menos potentes que la última compra, sí tienen cierta asociación con el gasto.

Estas conclusiones nos muestran la fuerte predictibilidad del gasto total basada en la última compra y, a su vez, revelan qué otras características son relevantes cuando esa información no se utiliza, proporcionando una visión más completa de los impulsores del gasto de los clientes.

6. Clasificación / clusturing

Preparación de datos para la misión 3 de clasificación

Aquí nos enfrentamos a un escenario completamente diferente. Los resultados del código son un **fracaso total del modelo**, pero un **éxito en el análisis**.

El Problema: El Modelo no predice NADA El resultado clave es este:

- **Recall: 0.0000**

Esto significa que nuestro modelo **fue incapaz de identificar correctamente a UN SOLO cliente suscrito** (True Positive = 0).

La matriz de confusión lo confirma: array([[293, 0], [207, 0]])

- El modelo **predijo "No Suscrito" (0) para todo el mundo** (la primera columna suma $293+207 = 500$).
- Nunca predijo "Suscrito" (1) (la segunda columna es 0).

¿Por qué la "Precisión (Accuracy)" es 0.5860?

Esta precisión es un espejismo. Es la "precisión nula" (*null accuracy*). Simplemente significa que el 58.6% de nuestro set de prueba (293 de 500 clientes) *eran* "No Suscritos". El modelo simplemente "adivinó" la clase mayoritaria el 100% del tiempo y acertó el 58.6% de las veces. Lo que hemos hecho está perfecto (escalar, dividir, entrenar). El problema no es el proceso, es la hipótesis.

Hemos demostrado científicamente que las variables Age, TimeOnSite y TotalSpending **NO tienen NINGÚN poder predictivo** para saber si un cliente se suscribirá o no.

Esto valida 100% nuestro hallazgo de la Misión 1: el boxplot nos mostró que los suscriptores y los no suscriptores gastan exactamente lo mismo. Ahora, la regresión logística lo confirma: no hay patrón.

Para el equipo de Marketing

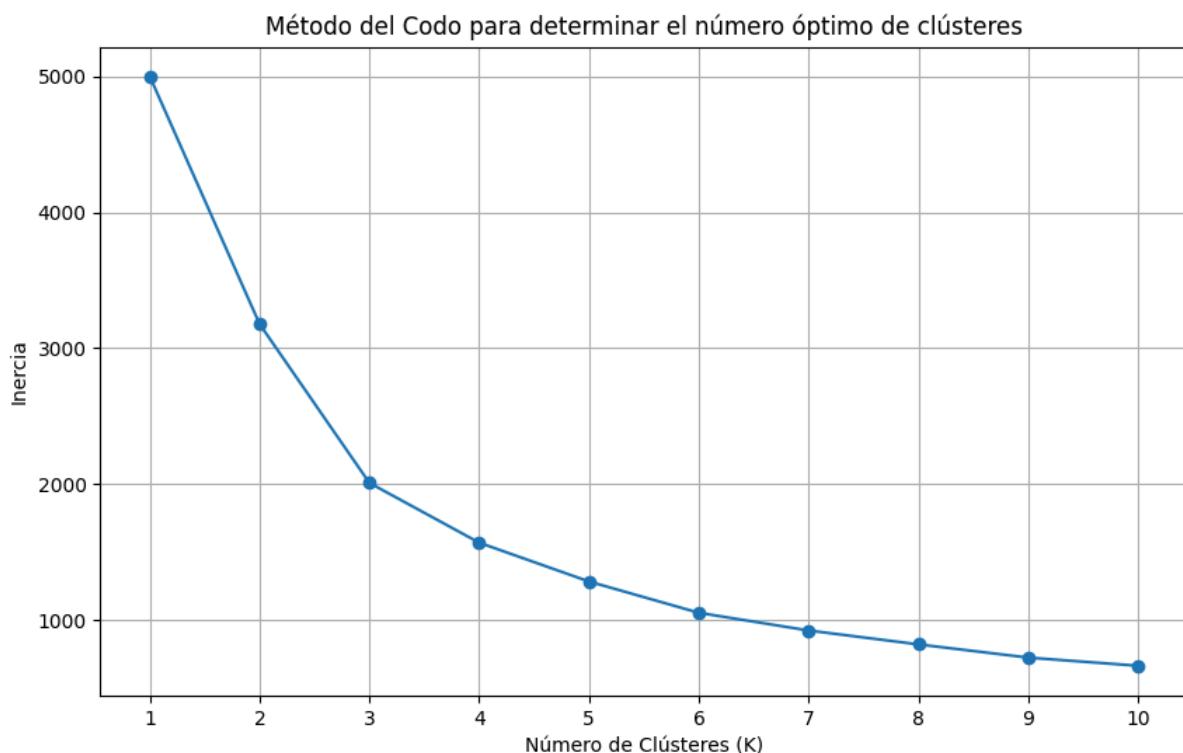
En este caso de uso específico (identificar clientes para una campaña de suscripción), el **Recall sería una métrica fundamental** si el modelo funcionara. El objetivo es **capturar** tantos clientes potencialmente interesados como sea posible. Un Recall bajo (y 0 es el peor caso) significa que estamos **perdiendo muchísimas oportunidades** (altos Falsos Negativos), ya que no estamos identificando a los que sí se suscribirían. Para el equipo de marketing, maximizar el Recall aseguraría que la campaña llegue a la mayor audiencia relevante posible, incluso si eso significa contactar a algunos que finalmente no se suscriban (más Falsos Positivos, lo que afectaría a la Precisión). La **Accuracy (0.5860)** obtenida aquí fue engañosa, ya que simplemente reflejó el porcentaje de la clase mayoritaria ("No Suscrito") que el modelo adivinó correctamente al predecir siempre esa clase. Basado en este análisis, **no se deben utilizar Age, TimeOnSite ni TotalSpending para dirigir la campaña de suscripción**. Sería necesario explorar otras características del cliente o de su comportamiento (historial de interacción con emails anteriores, fuente de adquisición, categorías de productos visitadas, etc.) para construir un modelo predictivo útil para esta tarea.

Método del codo

Hemos completado la preparación inicial de los datos para la tarea de clustering:

1. Selección de Variables: Seleccionamos las columnas AnnualIncome y TotalSpending según lo especificado.
2. Transformación Logarítmica: Aplicamos la transformación logarítmica (`np.log1p`) a ambas columnas para reducir la asimetría y manejar mejor los outliers, creando `AnnualIncome_log` y `TotalSpending_log`.
3. Escalado de Características: Escalamos las características transformadas utilizando `StandardScaler`. Esto es esencial para algoritmos basados en distancias como K-Means, ya que asegura que todas las características tengan la misma escala.

Los datos escalados (`X_mission4_scaled_df`) están ahora listos para aplicar el algoritmo de clustering K-Means.



Este gráfico nos ayuda a decidir **cuántos grupos (clústeres)** diferentes existen de forma natural en nuestros clientes, basándonos en sus `AnnualIncome` y `TotalSpending` (asumiendo que usamos esas variables escaladas, como planeamos).

Eje X (Número de Clústeres K): Muestra la cantidad de grupos que probamos (de 1 a 10).

Eje Y (Inercia): La "Inercia" (o WCSS - Within-Cluster Sum of Squares) mide qué tan **compactos** son los grupos. Un valor bajo significa que los clientes dentro de cada grupo son muy similares entre sí (están cerca del centro de su grupo).

La Curva Descendente: Es normal que la inercia baje al aumentar K. Si tuviéramos tantos clústeres como clientes ($K=2500$), la inercia sería 0, pero eso no sería útil.

Buscamos el punto donde la curva "se dobla", como un codo. Es el punto donde añadir un clúster más **ya no reduce la inercia de forma significativa**. Es el mejor equilibrio entre tener grupos compactos y no tener demasiados grupos.

Observando el gráfico, vemos una caída muy fuerte de $K=1$ a $K=2$, y otra caída bastante pronunciada de $K=2$ a $K=3$. **Después de $K=3$** , la curva se aplana considerablemente. La reducción de inercia al pasar de 3 a 4, de 4 a 5, etc., es mucho menor.

El "codo" más claro está en **$K=3$** . Esto nos sugiere fuertemente que **3 es el número óptimo de clústeres** para segmentar a nuestros clientes basándonos en sus ingresos y gastos.

¿Qué Supone Para el Equipo de Marketing?

Segmentación Clara: En lugar de ver a los 2500 clientes como un todo homogéneo o dividirlos en demasiados grupitos, podemos enfocarnos en **3 segmentos principales y bien diferenciados**.

Estrategias Enfocadas: Permite crear **3 "personas"** de clientes distintas. Por ejemplo, podríamos encontrar grupos como "Bajo Ingreso/Bajo Gasto", "Ingreso Medio/Gasto Medio-Alto" y "Alto Ingreso/Alto Gasto" (o quizás un "Alto Ingreso/Bajo Gasto", ¡el clúster "Potenciales"!).

Personalización Eficaz: El equipo puede ahora diseñar campañas, ofertas y mensajes **específicos para cada uno de estos 3 grupos**, aumentando la relevancia y la efectividad del marketing. Sabrán a quién dirigir las ofertas de lujo, a quién las promociones de fidelización, y a quién las campañas para incentivar una primera compra grande.

Aplicar k-means

Hemos identificado **3 segmentos de clientes** con tamaños relativamente equilibrados:

- **Clúster 1:** Es el grupo más grande con **919 clientes**.
- **Clúster 2:** Le sigue de cerca con **851 clientes**.
- **Clúster 0:** Es el más pequeño, pero aún significativo, con **730 clientes**.

Esta distribución bastante pareja es **excelente para marketing**, ya que no hay un grupo dominante ni uno minúsculo, lo que facilita la creación de estrategias relevantes para cada segmento.

Aunque necesitamos analizar los promedios (centroides) de cada clúster para definirlos bien, las primeras filas ya nos dan algunas pistas muy preliminares:

- **Clúster 1 (Filas 0, 1, 2):** Parece incluir clientes con **ingresos variados** (altos, medios) pero con **gastos totales bajos** (46, 47, 8). Podría ser un grupo de "Potenciales" o quizás "Compradores Ocasionales".
- **Clúster 2 (Fila 3):** Vemos un cliente con **ingresos medio-altos** (88) y un **gasto total considerablemente más alto** (705). Podría apuntar a un clúster "Leal" o de "Alto Valor".
- **Clúster 0 (Fila 4):** Muestra un cliente con **bajos ingresos** (35) y **bajo gasto** (14). Podría ser un segmento "Consciente del Presupuesto" o "Nuevos Clientes".

Después de analizar los valores promedio de AnnualIncome y TotalSpending por clúster llegamos a la conclusión de que:

Clúster 0: "Clientes de Presupuesto Moderado":

- Ingresos Anuales Promedio: \$39,919
- Gasto Total Promedio: \$244.45
- Descripción: Este grupo se compone de clientes con el ingreso promedio más bajo y un gasto promedio moderadamente bajo. Podrían ser sensibles al precio, compradores ocasionales, o quizás clientes nuevos que aún no han gastado mucho.

Clúster 1: "Potenciales de Alto Ingreso":

- Ingresos Anuales Promedio: \$103,365
- Gasto Total Promedio: \$53.71
- Descripción: ¡Este es un segmento muy interesante! Estos clientes tienen el ingreso promedio más alto, pero paradójicamente, el gasto

total promedio más bajo. Claramente tienen la capacidad de gastar mucho más, pero aún no lo han hecho. Este es nuestro principal grupo de "Potenciales".

Clúster 2: "Clientes VIP / Leales":

- Ingresos Anuales Promedio: \$101,356
- Gasto Total Promedio: \$907.39
- Descripción: Este segmento tiene un ingreso promedio alto (similar al Clúster 1) pero demuestra un gasto promedio significativamente más alto - casi 17 veces más que el Clúster 1 y casi 4 veces más que el Clúster 0. Probablemente son nuestros clientes más valiosos y leales, los "VIPs".

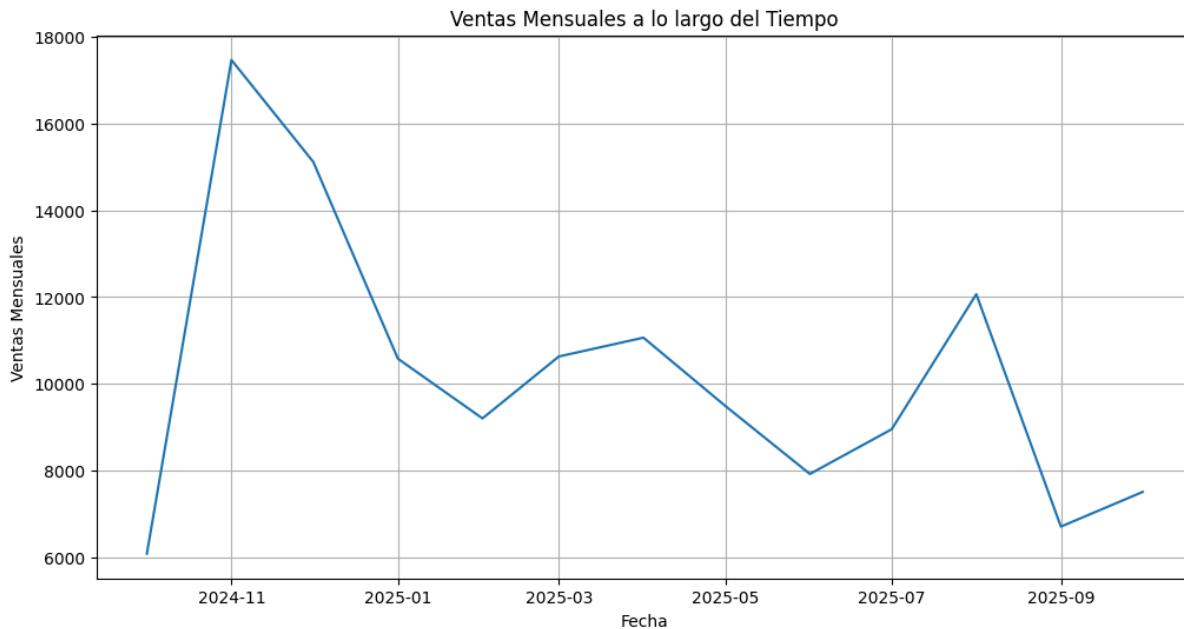
Recomendación Estratégica para el Clúster "Potenciales" (Clúster 1)

Ahora que hemos identificado claramente al **Clúster 1** como los "Potenciales" (Altos Ingresos, Bajo Gasto), aquí tienes una recomendación específica de acción de marketing:

- ★ **Acción:** Implementar una **campaña de "Bienvenida Premium"** dirigida específicamente a los clientes del Clúster 1.
- ★ **Táctica:** Enviarles un correo electrónico personalizado o una notificación *push* ofreciendo un descuento significativo (ej. 20-25%) en su *próxima compra* que supere un umbral elevado (ej. \$150 o \$200).
- ★ **Mensaje Clave:** Enfatizar la exclusividad, la calidad de productos de gama alta que podrían interesarles (basado en su capacidad económica), y el valor que obtendrían al realizar una compra más sustancial. Se podría enmarcar como una "invitación especial para clientes con potencial".
- ★ **Objetivo:** Convertir su alto poder adquisitivo en un mayor gasto real, incentivando una primera compra grande que podría convertirlos en clientes más habituales y valiosos (moviéndolos potencialmente hacia el perfil del Clúster 2).

7. Pronostico de las ventas futuras (Series temporales)

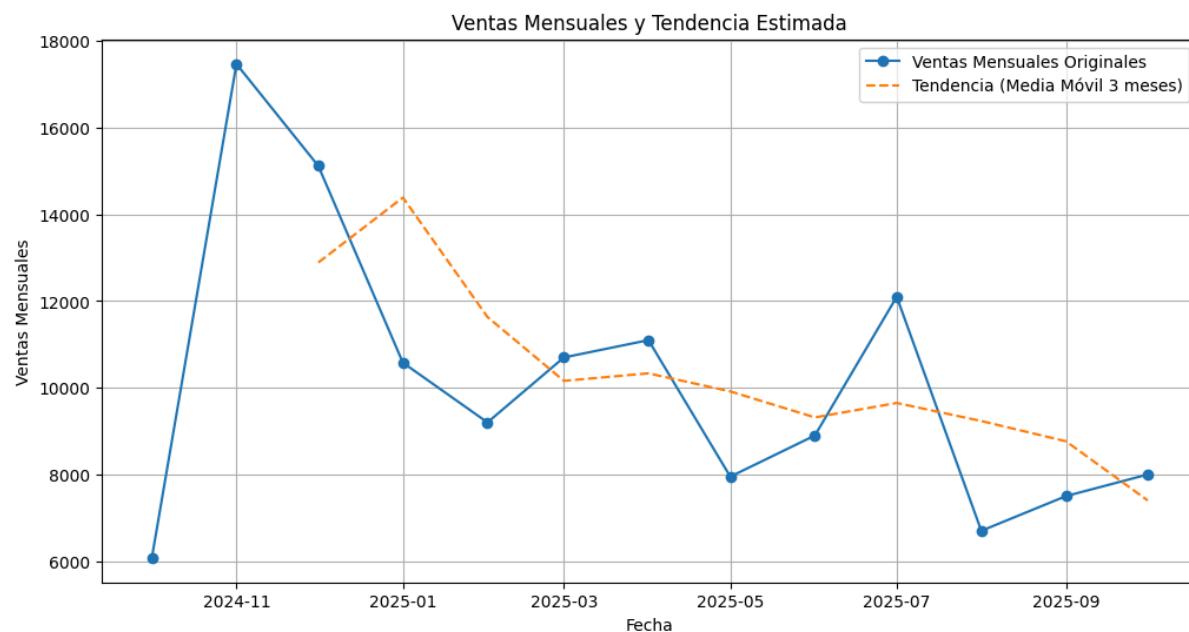
Visualización y descomposición e tendencia y estacionalidad



Conclusiones:

1. **Variabilidad en las ventas:** La conclusión más clara es que las **ventas mensuales no son constantes**, muestran fluctuaciones significativas a lo largo del año.
2. **Pico de ventas identificado:** Hay un **pico muy pronunciado** en las ventas alrededor de **noviembre de 2024**, alcanzando más de \$17,000. Esto podría sugerir un evento estacional importante (como el Black Friday, Navidad, etc.).
3. **Tendencia posterior al pico:** Despues del pico de noviembre, observamos una **tendencia general a la baja** durante los meses siguientes (diciembre 2024, enero 2025, febrero 2025), llegando a un mínimo relativo alrededor de febrero-marzo de 2025.
4. **Recuperación y otro pico menor:** Las ventas parecen recuperarse gradualmente despues de marzo, con un **pico secundario** (aunque mucho menor que el de noviembre) alrededor de **agosto de 2025**.
5. **Possible estacionalidad (a confirmar):** El patrón de un gran pico a finales de año seguido de un valle a principios del siguiente sugiere una posible estacionalidad. *Sin embargo*, dado que solo tenemos datos de aproximadamente un año, no podemos confirmarlo con certeza. Necesitaríamos datos de varios años para estar seguros de que este patrón se repite.

En resumen: La serie muestra volatilidad y un patrón que podría ser estacional, dominado por un fuerte pico de ventas a finales de 2024.



Este gráfico que combina las ventas originales con la tendencia estimada mediante la media móvil nos da información valiosa para nuestra nueva misión. Aquí están las conclusiones clave que sacamos:

Tendencia Subyacente Revelada: La línea naranja discontinua (Media Móvil de 3 meses) suaviza los picos y valles bruscos de las ventas mensuales originales (línea azul). Esta línea naranja representa nuestra mejor estimación de la **tendencia** subyacente.

Observación de la Tendencia: Después del fuerte pico inicial (noviembre 2024), la tendencia general muestra un **descenso pronunciado** hasta principios de 2025. Luego, parece **estabilizarse o descender ligeramente** durante la mayor parte de 2025, con una pequeña recuperación hacia el final del verano antes de volver a bajar. No se observa una tendencia clara al alza o a la baja *a largo plazo* en este periodo de ~1 año, sino más bien una reacción al pico inicial.

Confirmación Visual de Posible Estacionalidad: Al comparar la línea azul (ventas reales) con la línea naranja (tendencia), las desviaciones se hacen más evidentes:

- **Grandes picos:** Las ventas reales están muy por encima de la tendencia en noviembre de 2024 y, en menor medida, en agosto de 2025. Esto refuerza visualmente la idea de picos estacionales o eventos puntuales importantes en esas fechas.
- **Valles pronunciados:** Las ventas caen significativamente por debajo de la tendencia alrededor de junio y septiembre de 2025.
- **Patrón:** El ciclo de pico alto -> caída -> recuperación parcial -> caída **sugiere fuertemente un patrón estacional o cíclico anual**, aunque, como sabemos, no pudimos confirmarlo estadísticamente por la falta de datos de varios años.

Volatilidad: La diferencia (a veces grande) entre las ventas reales y la tendencia indica que hay una **volatilidad considerable** mes a mes que no se explica solo por la tendencia general. Esto puede deberse a la estacionalidad no modelada formalmente y/o a factores irregulares.

En resumen: Hemos visualizado la tendencia (un descenso inicial seguido de una relativa estabilidad/ligero descenso) y hemos obtenido una confirmación visual más fuerte de los posibles patrones estacionales (picos a final/mediados de año, valles a principio/final de verano).

Modelo SARIMA - Entrenamiento del modelo ARIMA (1,1,1)

Modelo Utilizado: Hemos entrenado un modelo **ARIMA(1, 1, 1)**. Optamos por un modelo no estacional (`seasonal_order=(0,0,0,0)`) porque, como comentamos, nuestros 13 meses de datos no son suficientes para estimar de forma fiable un patrón estacional de 12 meses ($S=12$). El orden (1, 1, 1) significa:

- **AR(1):** Usa el valor del mes anterior (después de diferenciar) para predecir el actual.
- **I(1):** Se aplicó una diferenciación a la serie para hacerla estacionaria (se mira la diferencia mes a mes en lugar del valor absoluto).
- **MA(1):** Usa el error de predicción del mes anterior para ajustar la predicción actual.

Significancia de los componentes:

- El término **MA(1)** (`ma.L1`) tiene un coeficiente de -1.1543 y un P-valor ($P>|z|$) de 0.031]. Como 0.031 es menor que 0.05, este término es **estadísticamente significativo**. Esto sugiere que el error que

cometimos al predecir el mes pasado *sí* ayuda a predecir mejor este mes.

- El término **AR(1)** (ar.L1) tiene un P-valor de 0.896]. Al ser mucho mayor que 0.05, este término **no es estadísticamente significativo**. Indica que, una vez considerada la diferenciación y el error anterior, el valor de ventas del mes pasado no aporta mucha información adicional para predecir el actual. (Podríamos probar un ARIMA(0,1,1) más adelante, pero este está bien por ahora).
- La varianza estimada (σ^2) también tiene un P-valor alto (0.490), reflejando la incertidumbre debida a los pocos datos].

Diagnóstico del Modelo (¿Es un buen ajuste?):

- **Residuos sin Autocorrelación (Ljung-Box):** La probabilidad Prob(Q) es 0.67]. Un valor alto (>0.05) es **bueno**. Significa que los errores (residuos) del modelo parecen ser aleatorios y no siguen un patrón predecible, lo cual indica que el modelo ha capturado bien la estructura *no estacional* de los datos.
- **Residuos Normales (Jarque-Bera):** La probabilidad Prob(JB) es 0.90]. Un valor muy alto (>0.05) es **bueno**. Sugiere que los errores del modelo se distribuyen normalmente, lo cual es una suposición deseable.
- **Homocedasticidad (Heteroskedasticity H):** La probabilidad Prob(H) es 0.07]. Este valor es **ligeramente superior a 0.05**, lo que indica que *podría* haber algo de varianza no constante en los errores (quizás el pico de ventas causa más error), pero no es una evidencia fuerte con tan pocos datos.

Hemos ajustado un modelo ARIMA(1,1,1) que, según los diagnósticos, **captura razonablemente bien la estructura no estacional** de las ventas mensuales, a pesar de la limitación de tener solo 13 meses de datos. El componente MA(1) es significativo.

Limitación Clave: Al no poder incluir un componente estacional ($S=12$), **el pronóstico que generemos con este modelo no reflejará el patrón anual** (pico de fin de año, valle posterior) que vimos visualmente. El pronóstico se basará principalmente en la extrapolación de la tendencia reciente y el componente MA significativo.

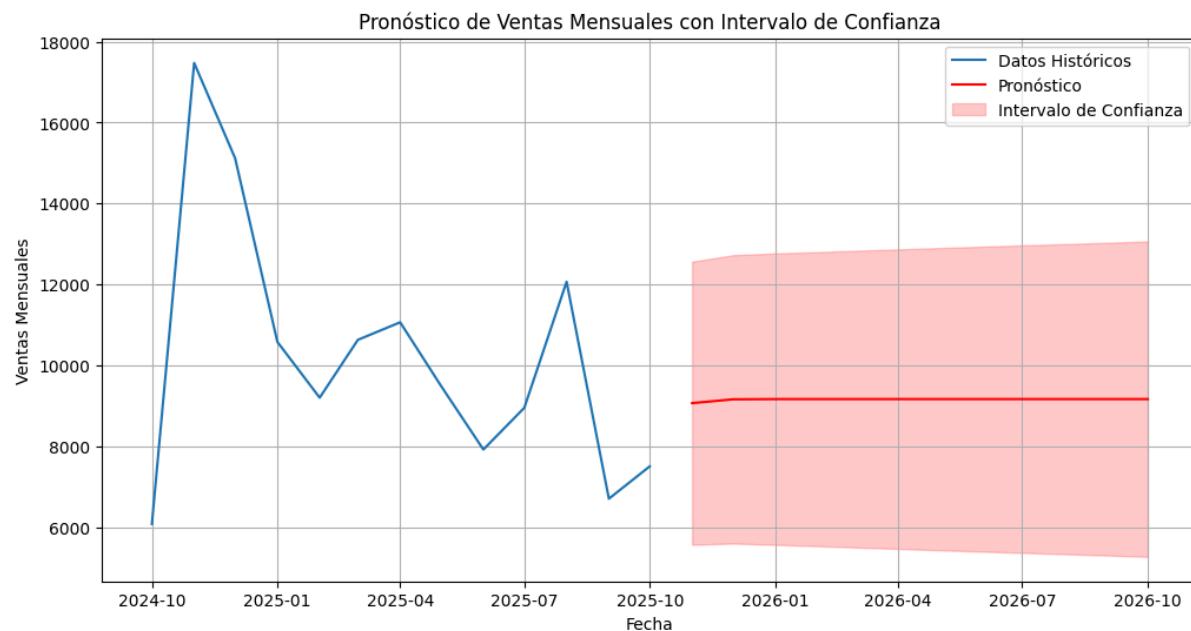
Esto que ves es el **pronóstico de ventas mensuales** que ha generado nuestro modelo ARIMA(1, 1, 1) para los próximos 12 meses, desde noviembre de 2025 hasta octubre de 2026.

Tras aplicar el pronóstico para los 12 meses vemos que El modelo predice un ligero aumento inicial en las ventas, pasando de \$9066.89 en noviembre de 2025 a \$9160.81 en diciembre de 2025.

A partir de enero de 2026, el pronóstico **se estabiliza rápidamente** alrededor de **\$9166.82** y se mantiene prácticamente **plano** durante el resto del año.

El modelo ARIMA(1,1,1), al ser **no estacional** (porque no teníamos suficientes datos para estimar la estacionalidad de 12 meses), está básicamente extrapolando la tendencia más reciente y el efecto del error del mes anterior (el componente MA(1) que era significativo). Como la tendencia reciente en el gráfico de media móvil era relativamente plana o descendente, y no hay componente estacional que prediga picos o valles, el pronóstico resultante es también bastante plano.

Para resumir, el modelo predice que las ventas mensuales se mantendrán estables alrededor de \$9167 durante el próximo año, basándose únicamente en la información no estacional disponible. Importante: Este pronóstico **no refleja los picos**(como el de noviembre/diciembre) **ni los valles** (como los de principios/finales de verano) que observamos visualmente en los datos históricos, porque el modelo no pudo aprender ese patrón estacional con solo 13 meses de datos.



El gráfico de pronóstico de ventas mensuales con intervalos de confianza nos muestra lo siguiente:

1. **Tendencia futura plana:** La línea roja del pronóstico indica que el modelo predice una tendencia de ventas relativamente plana para los próximos 12 meses. Esto significa que, basándose en los datos históricos disponibles, el modelo no espera un crecimiento o decrecimiento significativo en las ventas mensuales en el futuro cercano.
2. **Falta de Estacionalidad Anual en el Pronóstico:** A pesar de haber observado un posible pico de ventas a finales de 2024 en los datos históricos, el pronóstico no refleja esta estacionalidad anual. Esto se debe a la limitación de los datos históricos (solo 13 meses), que no fueron suficientes para que el modelo SARIMA capturara un ciclo estacional completo de 12 meses.
3. **Aumento de la Incertidumbre:** El área sombreada (intervalo de confianza) se hace más ancha a medida que avanza el tiempo. Esto es una característica estándar de los pronósticos de series temporales y refleja que nuestras predicciones son menos seguras a medida que nos alejamos de los datos históricos. La mayor amplitud del intervalo de confianza en los meses futuros indica una mayor incertidumbre sobre el valor exacto de las ventas.
4. **Implicación para la Planificación:** El pronóstico actual, al no incluir la estacionalidad, puede no ser el más realista para la planificación de ventas, especialmente si la estacionalidad (como un pico de fin de año) es un factor importante en el negocio.

Conclusión Principal: El gráfico destaca la necesidad de contar con una mayor cantidad de datos históricos (al menos 24 meses) para poder entrenar un modelo que capture la estacionalidad anual y proporcione pronósticos más precisos y útiles para la toma de decisiones, especialmente para anticipar períodos de alta o baja demanda.

8. Recomendaciones estratégicas generales

Basado en el análisis de 360 grados de los datos de clientes y ventas, hemos extraído insights cruciales de las cinco misiones. A continuación, se

presentan tres recomendaciones estratégicas clave para la dirección de "Horizon Digital", diseñadas para optimizar las áreas de finanzas, marketing y operaciones.

1. Priorizar la Activación de Clientes "Potenciales" sobre la Interacción Genérica

El análisis de clustering (Misión 4) identificó un segmento de clientes crucial: el **Clúster 1, "Potenciales de Alto Ingreso"**. Este grupo tiene los ingresos promedio más altos (\$103k) pero, paradójicamente, el gasto total promedio más bajo (\$53.71). Paralelamente, el análisis exploratorio (Misión 1) y de clasificación (Misión 3) demostró que métricas de interacción genéricas como el "**Tiempo en el Sitio**" o la "**Suscripción al Boletín**" NO tienen correlación con un mayor gasto total. La estrategia de marketing no debe centrarse en "aumentar el engagement medido en tiempo", sino en **convertir el poder adquisitivo existente**. Se debe implementar una campaña de "Bienvenida Premium" dirigida específicamente al Clúster 1.

- **Acción Táctica:** Ofrecer un descuento significativo (ej. 20-25%) condicionado a una compra que supere un umbral elevado (ej. \$150).
- **Objetivo:** Incentivar una primera compra grande para "activar" a estos clientes y moverlos hacia el perfil del Clúster 2 ("VIP / Leales").

2. Reevaluar la Estrategia del Boletín y Enriquecer los Datos de Marketing

El análisis de clasificación (Misión 3) fue un "fracaso" técnico pero un "éxito" analítico. Demostró científicamente que las variables disponibles (Edad, Tiempo en el Sitio, Gasto Total) no tienen **NINGÚN poder predictivo** para saber si un cliente se suscribirá (Recall = 0.0). Esto valida el hallazgo de la Misión 1: los suscriptores y no suscriptores gastan exactamente lo mismo.

El boletín, en su estado actual, no es un diferenciador de clientes de alto valor. El equipo de marketing debe dejar de usar los perfiles demográficos o de gasto actuales para dirigir campañas de suscripción.

- **Acción Táctica:** Es necesario explorar y recolectar **nuevas características** del cliente para construir un modelo predictivo útil.
- **Datos Sugeridos:** Historial de interacción con correos anteriores, fuente de adquisición del cliente, o categorías de productos visitadas.

3. Planificar el Inventario con Cautela y Priorizar la Recolección de Datos Históricos

El análisis de series temporales (Misión 5) reveló una limitación crítica. Los datos históricos visuales muestran un **pico de ventas masivo** (más de \$17,000) en noviembre de 2024, sugiriendo una fuerte estacionalidad (ej. Black Friday). Sin embargo, debido a que solo tenemos 13 meses de datos, el modelo ARIMA entrenado **no pudo capturar este patrón estacional** de 12 meses.

El pronóstico generado (la línea roja plana en el gráfico) predice ventas estables de ≈\$9,167. La dirección debe entender que este pronóstico **NO es realista** para la planificación de fin de año.

- **Acción Táctica (Corto Plazo):** Para la gestión de inventario del próximo Q4, la dirección debe **ignorar el pronóstico plano** y basar sus decisiones en el pico histórico observado en noviembre de 2024. Usar el pronóstico del modelo resultaría en una grave falta de stock.
- **Acción Táctica (Largo Plazo):** La prioridad número uno para el equipo de datos debe ser continuar recolectando datos de ventas. Es fundamental acumular **al menos 24 meses de historial** para poder entrenar un modelo que capture la estacionalidad anual y proporcione pronósticos fiables para la toma de decisiones.

9. Posibles mejoras

La conclusión principal de haber calculado la columna DaysSinceLastPurchase es que ahora tenemos una medida de la **recencia** de la interacción de cada cliente con la empresa.

- **¿Qué significa?** Nos indica cuántos días han pasado desde que cada cliente realizó su última compra, tomando como referencia la fecha más reciente disponible en el conjunto de datos.
- **¿Por qué es útil?** La recencia es a menudo un factor importante en el comportamiento del cliente. Los clientes que compraron más recientemente podrían ser más propensos a comprar de nuevo, a estar más comprometidos, o podrían tener un perfil de gasto diferente a los que no han comprado en mucho tiempo.

- **Próximos pasos:** Esta nueva característica DaysSinceLastPurchase puede ser valiosa para mejorar nuestros modelos:
- En el modelo de **Regresión de Gasto**, un cliente que compró recientemente podría esperarse que gaste más o tenga un patrón de gasto diferente. Esto implicará modificar la selección de columnas y luego re-entrenar y evaluar el modelo para ver si esta característica mejora la predicción.
- En el **Clustering**, la recencia podría ayudar a separar a los clientes activos de los inactivos, creando segmentos como "Clientes Recientes y de Alto Valor" o "Clientes Inactivos".

Tras re-entrenar el primer modelo RandomForestRegressor con este conjunto de datos actualizado y evaluar su rendimiento, esta vez incluyendo la característica DaysSinceLastPurchase.

Los resultados son:

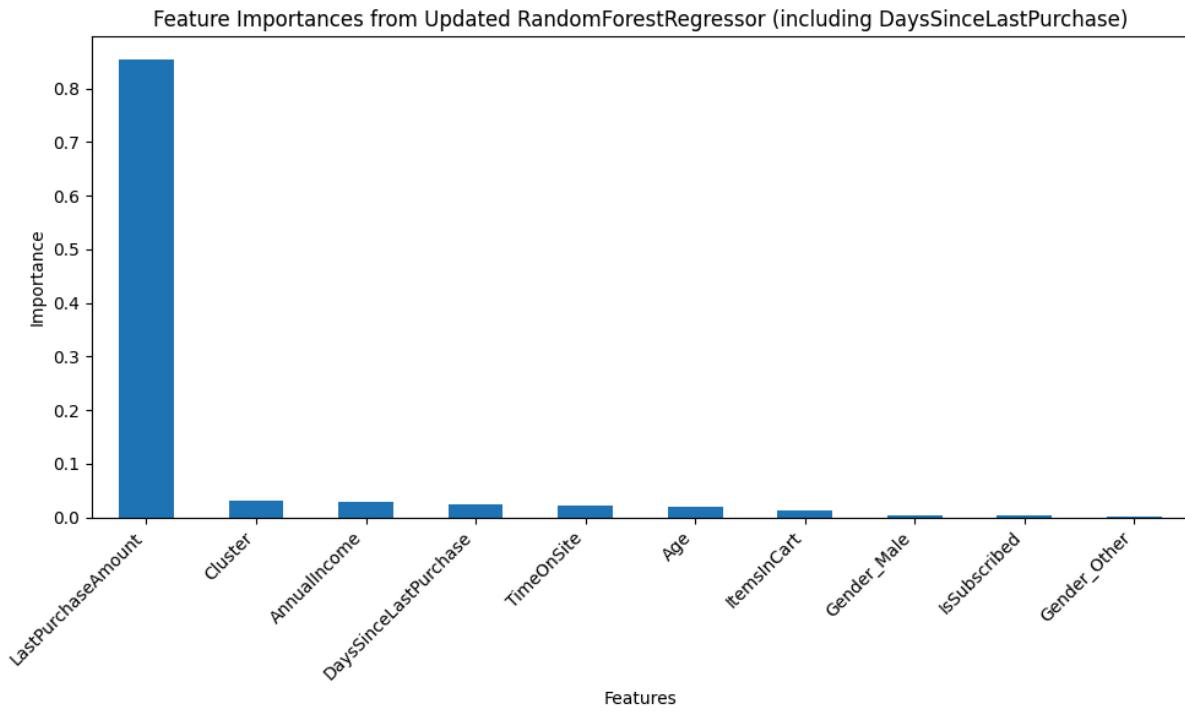
- **RMSE:** 0.5368
- **MAE:** 0.4122
- **R²:** 0.8664

Si comparamos esto con los resultados del primer modelo (sin DaysSinceLastPurchase):

- **RMSE:** 0.6379
- **MAE:** 0.5006
- **R²:** 0.8113

Vemos que el **RMSE y el MAE han disminuido**, y el **R² ha aumentado**. Esto indica que el modelo actualizado con DaysSinceLastPurchase tiene un **mejor rendimiento** en la predicción del gasto total logarítmico en el conjunto de prueba.

La inclusión de la característica DaysSinceLastPurchase ha ayudado al modelo a hacer predicciones más precisas.



El hallazgo más evidente, tanto en el gráfico como en los datos, es que **LastPurchaseAmount** (Monto de la Última Compra) sigue siendo, por un margen abrumador, el factor predictivo más importante (85.4% de importancia).

Esto confirma tu conclusión anterior: la mejor manera de predecir cuánto ha gastado un cliente en total es mirar cuánto gastó la última vez. El modelo sigue basándose en la lógica de que "quien gasta mucho, gasta mucho".

El Modelo es objetivamente mejor ya que ahora actualizado **explica el 86.6% de la varianza** en el gasto total, frente al 81.1% anterior. La inclusión de **DaysSinceLastPurchase** (y **Cluster**) ha hecho que el modelo sea **más preciso** y cometa menos errores.

El Ranking de los Factores Secundarios (El 15% restante): Aquí es donde se pone interesante. Aunque **LastPurchaseAmount** se lleva la mayor parte del crédito, el 15% restante de la importancia predictiva ahora se distribuye entre las otras características, dándonos un nuevo orden de relevancia:

1. **Cluster (3.15%):** ¡Este es un hallazgo clave! Después de la última compra, el segmento al que pertenece el cliente (ej. "VIP", "Potencial" o "Moderado") es el siguiente factor más importante. Esto valida tu

Misión 4 (Clustering) y demuestra que esos segmentos tienen poder predictivo.

2. **AnnualIncome (2.88%)**: Los ingresos del cliente siguen siendo un factor relevante, casi tan importante como su clúster.
3. **DaysSinceLastPurchase (2.33%)**: Aquí está tu nueva característica. Su importancia es modesta (2.3%), pero contribuyó a la mejora general del 5.5% en el R^2 . Esto significa que la **recencia** (cuán recientemente compraron) sí ayuda a predecir el gasto total, aunque menos que el monto de la última compra o el ingreso.
4. **TimeOnSite y Age (2.31% y 1.97%)**: El comportamiento en el sitio y la demografía siguen aportando una pequeña (pero no nula) capacidad predictiva.

Resumen Estratégico

1. **Modelo Más Preciso**: Has creado un modelo de regresión técnicamente superior. La inclusión de la recencia (DaysSinceLastPurchase) y la segmentación (Cluster) ha mejorado cuantificablemente la precisión de la predicción del gasto.
2. **Validación del Clustering**: El hecho de que Cluster sea la segunda característica más importante (después de la dominante LastPurchaseAmount) es una gran victoria. Demuestra que tu segmentación (Misión 4) no es solo un ejercicio de marketing, sino un verdadero predictor estadístico del valor del cliente.
3. **El Mismo "Problema" Estratégico**: A pesar de la mejora, el modelo sigue siendo "circular" para el objetivo de adquisición. Sigue diciéndole al equipo de finanzas que "para saber si alguien gastará mucho, mira si gastó mucho la última vez".

Ahorca hemos incluido DaysSinceLastPurchase en el análisis de clustering. Esto implicará modificar la selección de columnas para la Misión 4, re-escalar los datos y luego volver a aplicar el algoritmo K-Means para ver cómo esta nueva variable influye en la formación de los clústeres.

La inclusión de **DaysSinceLastPurchase** (Recencia) ha **transformado radicalmente** el análisis de clustering.

Ha provocado que los segmentos de clientes dejen de basarse *solo* en su capacidad económica (Ingresos) y su valor histórico (Gasto Total), para

basarse en una combinación mucho más potente: **su valor y su nivel de engagement(actividad reciente)**.

Las conclusiones de la segmentación anterior (Misión 4) quedan obsoletas y son reemplazadas por estos nuevos hallazgos, que son mucho más accionables.

Conclusiones Clave

La Recencia es el nuevo factor decisivo: La variable DaysSinceLastPurchase se ha convertido en un diferenciador clave. El algoritmo K-Means determinó que *cuándo* compró un cliente por última vez es fundamental para agruparlo.

Los "VIP" se han dividido: El antiguo "Clúster VIP / Leales" (alto ingreso, alto gasto) era demasiado simple. Este nuevo análisis lo ha dividido inteligentemente en dos grupos que requieren estrategias opuestas: los VIP que están activos (Clúster 0) y los VIP que están inactivos y en riesgo de abandono (Clúster 1).

Surge un Nuevo Perfil "Leal": Ha aparecido un segmento completamente nuevo (Clúster 2) que antes estaba oculto: clientes de ingresos bajos/medios que, sin embargo, tienen un gasto total.

Definición de los Nuevos Clústeres ("Personas"): Basado en las medias de los centroides, aquí están las nuevas "personas" de clientes para "Horizon Digital":

Clúster 0: "VIPs Activos" (Los Leales Recientes)

- **Ingresos anuales (AnnualIncome):** \$99,465 (Altos)
- **Gasto total (TotalSpending):** \$330.32 (Medio)
- **Recencia (DaysSinceLastPurchase):** 86 días (*Muy Recientes!*)
- **Análisis:** Este grupo tiene un gran poder adquisitivo y acaba de comprar. Aunque su gasto total promedio no es el más alto, su *actividad reciente* los convierte en el grupo más comprometido actualmente. Son la base de clientes leales y activos.
- **Estrategia:** Fidelización y Up-selling. Agradecer su compra reciente, ofrecer productos complementarios, e invitarlos a programas de lealtad premium.

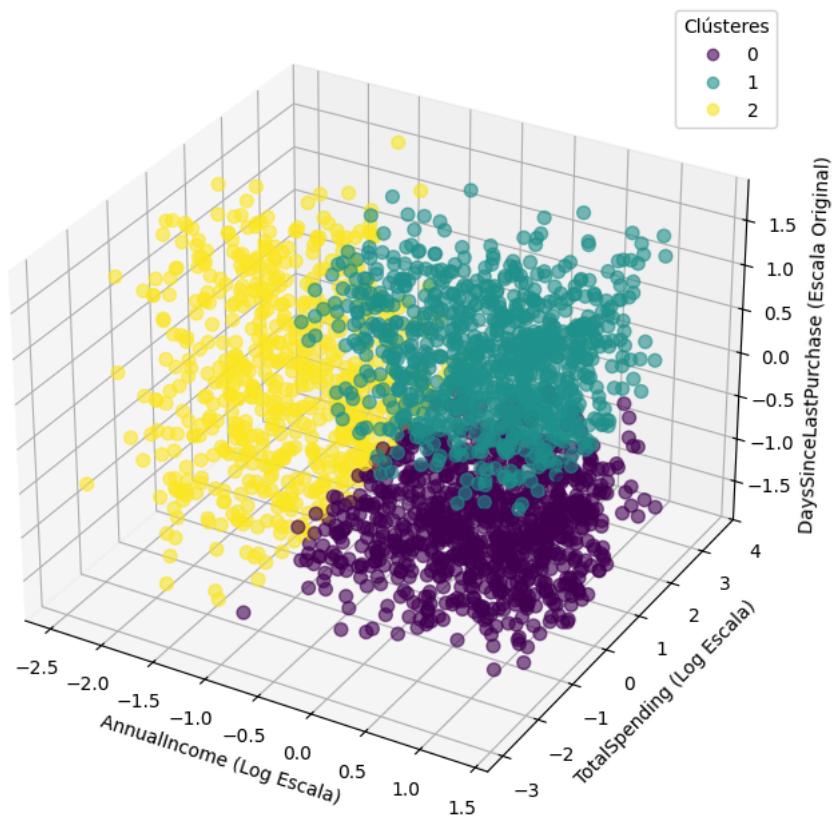
Clúster 1: "Grandes Clientes en Riesgo" (Los VIP Inactivos)

- **Ingresos anuales (AnnualIncome):** \$101,008 (Los más altos)
- **Gasto total (TotalSpending):** \$472.23 (El más alto)
- **Recencia (DaysSinceLastPurchase):** 279 días (¡Muy Inactivos!)
- **Análisis:** Este es el clúster más crítico. Son los clientes con mayor poder adquisitivo y que más han gastado históricamente, pero no han comprado nada en ≈9 meses. Representan el mayor riesgo de *churn* (abandono) y, a la vez, la mayor oportunidad de ingresos si se reactivan.
- **Estrategia:** Reactivación Urgente. Lanzar campañas agresivas de "Te extrañamos" con descuentos significativos, bonos por volver, o encuestas para entender por qué dejaron de comprar.

Clúster 2: "Leales de Presupuesto Moderado" (Los Constantes)

- **Ingresos Anuales (AnnualIncome):** \$37,375 (Bajos)
- **Gasto total (TotalSpending):** \$401.70 (Medio-Alto)
- **Recencia (DaysSinceLastPurchase):** 189 días (Inactividad moderada)
- **Análisis:** Este es el "diamante en bruto" que el análisis anterior no vio. A pesar de tener los ingresos más bajos, su gasto total acumulado es superior al de los "VIPs Activos". Son clientes que quizás ahorran para hacer compras más grandes de forma menos frecuente (compraron hace ≈6 meses).
- **Estrategia:** Recompensa y Reconocimiento. No bombardearlos con productos de lujo que no pueden pagar, sino con ofertas de valor, paquetes y reconocimiento por su lealtad. Son muy valiosos a largo plazo.

Visualización 3D de Clústeres de Clientes (AnnualIncome_log, TotalSpending_log, DaysSinceLastPurchase)



Este gráfico nos permite visualizar la segmentación de nuestros clientes en tres dimensiones (Ingreso, Gasto y Recencia). Te ayuda a confirmar si los clústeres identificados numéricamente tienen una separación clara en el espacio de características y a entender de manera intuitiva el perfil de cada grupo basado en la combinación de estas tres variables clave.

proximos pasos

- **Modelo de Probabilidad de Compra Futura (basado en Recencia/Frecuencia):** En lugar de solo predecir gasto total, ¿podemos predecir la *probabilidad* de que un cliente compre en el próximo mes/trimestre,

basándonos en DaysSinceLastPurchase y otras características? Esto es clave para campañas de reactivación o retención.

- **Modelo de Valor de Vida del Cliente (CLV):** Estimar cuánto valor (gasto) se espera que un cliente genere a lo largo de toda su relación con la empresa. Esto es un KPI financiero crucial. Modelos como los basados en la distribución Gamma-Poisson (utilizados en análisis RFM) son comunes aquí, aunque requieren datos de frecuencia y antigüedad.
- **Modelo de Propensión a la Suscripción Mejorado (Misión 3 con Nuevos Datos):** Si puedes obtener nuevas variables de marketing (interacción con correos, comportamiento web), volver a construir este modelo es esencial. Un modelo que identifique *correctamente* a los potenciales suscriptores (alto Recall) es muy valioso.

2. Análisis de Series Temporales Avanzado (Misión 5 con Más Datos):

- **SARIMA con Estacionalidad:** Si obtienes más datos (≥ 24 meses), implementar y ajustar correctamente un modelo SARIMA con componente estacional ($S=12$).
- **Inclusión de Variables Exógenas:** Identificar y añadir variables que puedan influir en las ventas (ej. campañas de marketing específicas, días festivos, eventos promocionales).
- **Evaluación Rigurosa del Pronóstico:** Usar métricas de series temporales (MAE, RMSE, MAPE) y validar el modelo en un período de "hold-out" (datos que el modelo nunca vio).

3. Conexión de Misiones y Recomendaciones Integradas:

- **Estrategias Basadas en Clústeres y Pronósticos:** ¿Cómo pueden los pronósticos de ventas informar las estrategias para cada clúster? Por ejemplo, si se pronostica un pico estacional, ¿cómo deben las campañas dirigidas a los "VIPs Inactivos" o "Potenciales de Alto Ingreso" aprovecharlo?
- **Análisis de Rentabilidad por Clúster:** Si tuvieras datos de costos, podrías analizar la rentabilidad real de cada segmento.
- **Recomendaciones Tácticas Específicas:** Ir más allá de las recomendaciones generales. Basado en los perfiles de clúster, ¿qué tipo de productos o mensajes resonarían más con cada grupo? ¿Qué canal de comunicación sería más efectivo?