escuela
británica de
artes creativas
y tecnología

Título de la tarea: Tarea M27-2.

Nombre del estudiante: Eliuth Misraim Rojas Villavicencio.

Tema: Proyecto de empresa aliada: entregable 3.

Institución educativa: Escuela Británica de Artes Creativas y Tecnología.

Introducción:

Descubriendo Oportunidades en el Mercado con Ciencia de Datos

En el competitivo mercado de productos de consumo, entender a fondo dónde y cómo una marca genera valor es fundamental para su crecimiento sostenible. Las decisiones estratégicas ya no pueden basarse únicamente en la intuición; requieren un respaldo sólido en los datos.

Este proyecto se sumerge en el universo de la marca **Vanish**, un referente en la categoría de cuidado de la ropa. A través de la aplicación de técnicas de **clustering y análisis de datos**, nuestro objetivo es transformar las cifras de ventas en una hoja de ruta estratégica.

Analizaremos datos de ventas, productos y regiones para responder preguntas clave:

- ¿Cuáles son los **segmentos de mercado** más importantes para Vanish?
- ¿En qué regiones y con qué productos la marca tiene un **desempeño sobresaliente**?
- ¿Dónde se encuentran las **áreas de oportunidad** o los puntos ciegos que podrían estar frenando el crecimiento?

Al finalizar, no solo tendremos una visualización clara de la estructura del mercado, sino que también contaremos con **insights accionables** para optimizar el portafolio de productos, afinar las estrategias de marketing y, en última instancia, fortalecer la posición de Vanish frente a la competencia.

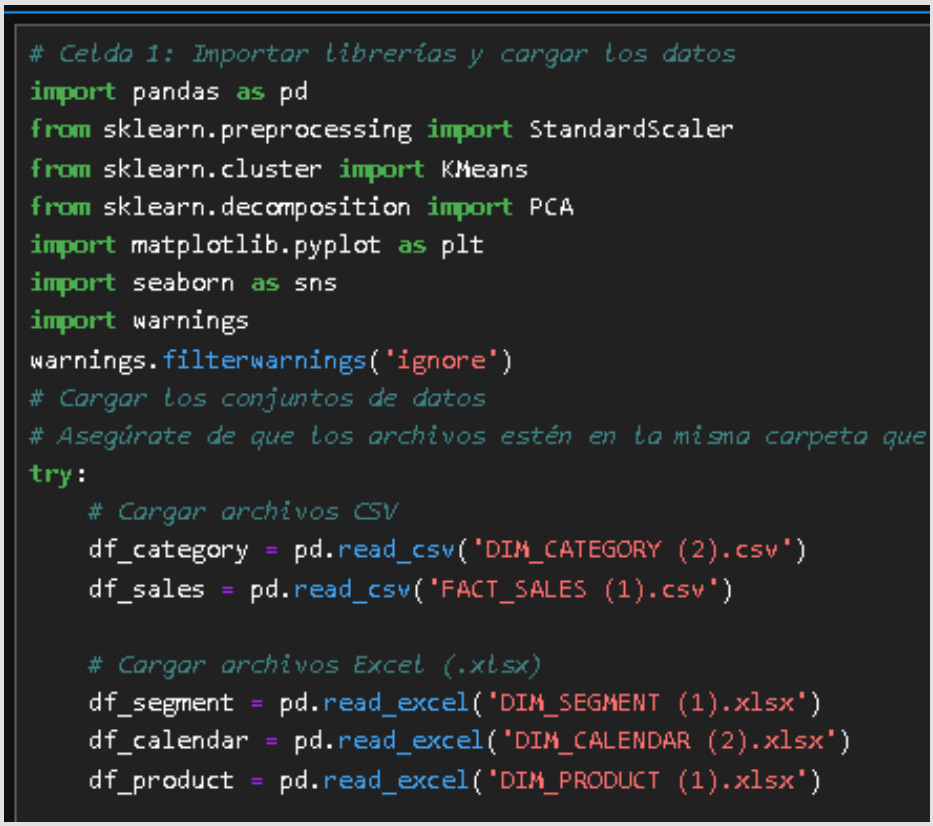
Metodología

1. **Consolidación de Datos:** Se unificaron múltiples archivos de ventas, productos y calendario en un único conjunto de datos, que luego se filtró para analizar exclusivamente la marca **Vanish**.
2. **Preparación para Clustering:** Se agruparon las ventas por segmento de mercado (región, categoría y formato). Posteriormente, las variables categóricas se convirtieron a formato numérico (One-Hot Encoding) y todas las características se estandarizaron para asegurar un análisis equitativo.
3. **Modelado y Segmentación:** Se utilizó el **Método del Codo** para determinar que **cuatro** era el número óptimo de clústers. A continuación, se aplicó el algoritmo **K-Means** para agrupar los segmentos de mercado en estos cuatro grupos distintos basándose en sus características.
4. **Análisis de Resultados:** Finalmente, se analizó la composición y el rendimiento de ventas de cada clúster para identificar patrones, fortalezas y áreas de oportunidad, traduciendo los resultados en insights estratégicos para la marca.

Metodología Detallada del Proyecto

1. **Consolidación de Datos:**

   - Se cargaron cinco fuentes de datos distintas (ventas, calendario, productos, categorías y segmentos) que se encontraban en formatos .xlsx y .csv.



```
# Celda 1: Importar librerías y cargar los datos
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
# Cargar los conjuntos de datos
# Asegúrate de que los archivos estén en la misma carpeta que
try:
    # Cargar archivos CSV
    df_category = pd.read_csv('DIM_CATEGORY (2).csv')
    df_sales = pd.read_csv('FACT_SALES (1).csv')

    # Cargar archivos Excel (.xlsx)
    df_segment = pd.read_excel('DIM_SEGMENT (1).xlsx')
    df_calendar = pd.read_excel('DIM_CALENDAR (2).xlsx')
    df_product = pd.read_excel('DIM_PRODUCT (1).xlsx')
```

   - Utilizando la librería Pandas, se fusionaron estas tablas en un único DataFrame maestro para crear una vista 360° del negocio.

```
# Fusionar ventas con calendario
df_sales_calendar = pd.merge(df_sales, df_calendar, on='WEEK', how='left')

# Fusionar productos con categorías
df_product.rename(columns={'CATEGORY': 'ID_CATEGORY'}, inplace=True)
df_product_category = pd.merge(df_product, df_category, on='ID_CATEGORY', how='left')

# Limpiar códigos de producto y fusionar ventas con productos
df_sales_calendar['ITEM_CODE_CLEAN'] = df_sales_calendar['ITEM_CODE'].str.replace('BP2', '', regex=False)
df_merged = pd.merge(df_sales_calendar, df_product_category, left_on='ITEM_CODE_CLEAN', right_on='ITEM', how='inner')

# Fusionar con segmentos
df_segment.rename(columns={'CATEGORY': 'ID_CATEGORY'}, inplace=True)
df_final = pd.merge(df_merged, df_segment, on=['ID_CATEGORY', 'ATTR1', 'ATTR2', 'ATTR3', 'FORMAT'], how='left')

# Seleccionar columnas relevantes
df_final = df_final[['YEAR', 'MONTH', 'REGION', 'MANUFACTURER', 'BRAND', 'ITEM_DESCRIPTION', 'CATEGORY', 'SEGMENT', 'TOTAL_VALUE_SALES']]
```

2. **Enfoque en la Marca de Interés:**

- Del DataFrame maestro, que contenía información de múltiples marcas, se aplicó un filtro para seleccionar únicamente los registros de ventas correspondientes a la marca Vanish.

```
# Filtrar solo por la marca 'VANISH' y crear una copia para evitar advertencias
df_vanish = df_final[df_final['BRAND'] == 'VANISH'].copy()

print("--- DataFrame de Vanish listo para el análisis ---")
df_vanish.head()
```

```
--- DataFrame de Vanish listo para el análisis ---
```

	YEAR	MONTH	REGION	MANUFACTURER	BRAND	ITEM_DESCRIPTION	CATEGORY	SEGMENT	TOTAL_VALUE_SALES
0	2022	8	TOTAL AUTOS AREA 5	RECKITT	VANISH	VANISH OXI ACTION GOLD QUITAMANCHAS BOLSA 1.8K...	FABRIC TREATMENT and SANIT\r\n	POWDER	116.519
1	2022	8	TOTAL AUTOS AREA 5	RECKITT	VANISH	VANISH OXI ACTION ROSA QUITAMANCHAS DOYPACK 24...	FABRIC TREATMENT and SANIT\r\n	POWDER	68.453
2	2022	8	TOTAL AUTOS AREA 5	RECKITT	VANISH	VANISH OXI ACTION GOLD QUITAMANCHA AHORRO DEL ...	FABRIC TREATMENT and SANIT\r\n	POWDER	1.481
3	2022	8	TOTAL AUTOS AREA 5	RECKITT	VANISH	VANISH INTELLIGENCE POLVO BOTE 450 GR NAL 7501...	FABRIC TREATMENT and SANIT\r\n	POWDER	182.839
4	2022	8	TOTAL AUTOS AREA 5	RECKITT	VANISH	VANISH OXI ACTION ROSA QUITA/MANCHAS P/ROPA BO...	FABRIC TREATMENT and SANIT\r\n	POWDER	0.679

3. **Definición de Segmentos de Mercado:**

- Para analizar el rendimiento a nivel de mercado y no por transacción individual, los datos se agruparon por tres variables clave: REGION, CATEGORY y SEGMENT.
- Para cada una de estas combinaciones, se calculó la suma total de ventas (TOTAL_VALUE_SALES), creando así los segmentos únicos que serían el foco del clustering.

```
# 1. Agrupar datos para la segmentación de mercados
df_segmentation = df_vanish.groupby(['REGION', 'CATEGORY', 'SEGMENT'])['TOTAL_VALUE_SALES'].sum().reset_index()
```

4. **Preparación de Características para el Modelo:**

- Codificación Numérica: Las variables de texto (categóricas) se convirtieron a un formato numérico usando la técnica *One-Hot Encoding*, indispensable para el algoritmo de clustering.

```
# 2. Convertir variables categóricas a numéricas (One-Hot Encoding)
df_encoded = pd.get_dummies(df_segmentation, columns=['REGION', 'CATEGORY', 'SEGMENT'])
```

- Estandarización: Todas las características, tanto las recién creadas como las ventas totales, se estandarizaron con StandardScaler. Este paso es crucial para asegurar que todas las variables tengan la misma importancia y escala en el análisis.

```
# 3. Estandarizar todas las características para el modelo
scaler = StandardScaler()
features_scaled = scaler.fit_transform(df_encoded)
```

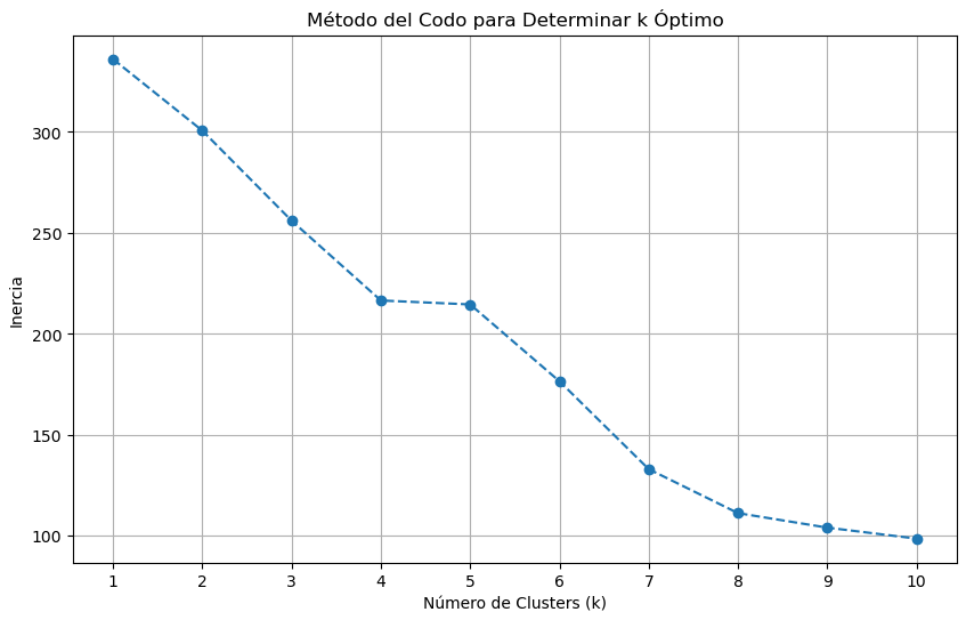
5. **Determinación del Número Óptimo de Clústers (k):**

- Se utilizó el Método del Codo, una técnica estándar para encontrar el número ideal de clústers. Se graficó el rendimiento del modelo para diferentes valores de k (del 1 al 10).

```
inertia = []
k_range = range(1, 11)

for k in k_range:
    kmeans = KMeans(n_clusters=k, random_state=42, n_init=10)
    kmeans.fit(features_scaled)
    inertia.append(kmeans.inertia_)
```

- La gráfica reveló un "codo" en k=4, indicando que este era el número de agrupaciones que ofrecía el mejor balance entre simplicidad y precisión.



6. **Aplicación del Modelo de Clustering:**

- Con k=4 como objetivo, se aplicó el algoritmo de clustering K-Means sobre los datos ya preparados.

```
# El codo se observa en k=4, por lo que será nuestro número de clusters
optimal_k = 4
kmeans = KMeans(n_clusters=optimal_k, random_state=42, n_init=10)
```

- El modelo asignó cada uno de los segmentos de mercado a uno de los cuatro clústers, basándose en la similitud de sus características.

```
# Asignar cada segmento a un cluster
df_segmentation['Cluster'] = kmeans.fit_predict(features_scaled)
```

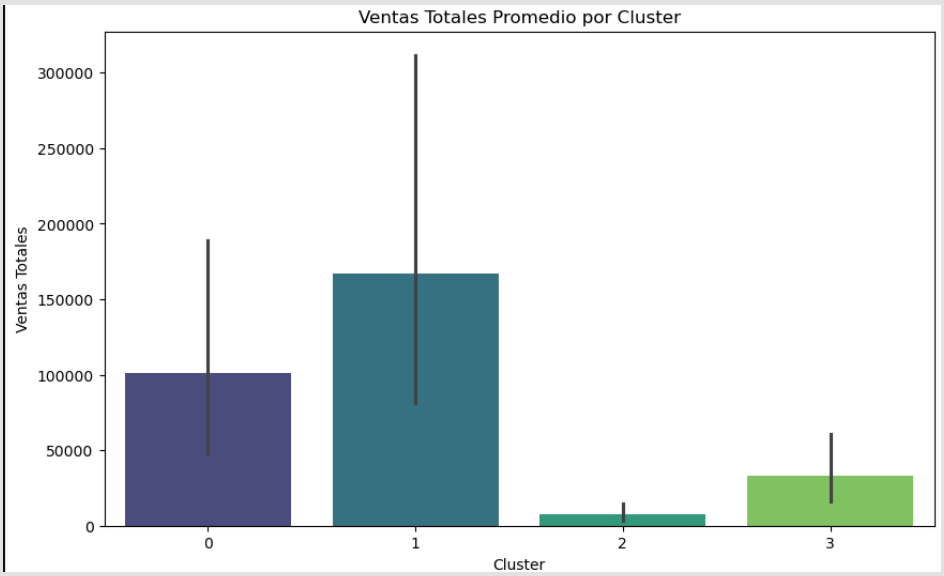
7. **Análisis y Visualización de Resultados:**

- Análisis Cuantitativo: Se calcularon estadísticas descriptivas (como la media de ventas) para cada uno de los cuatro clústers, lo que permitió cuantificar su rendimiento.

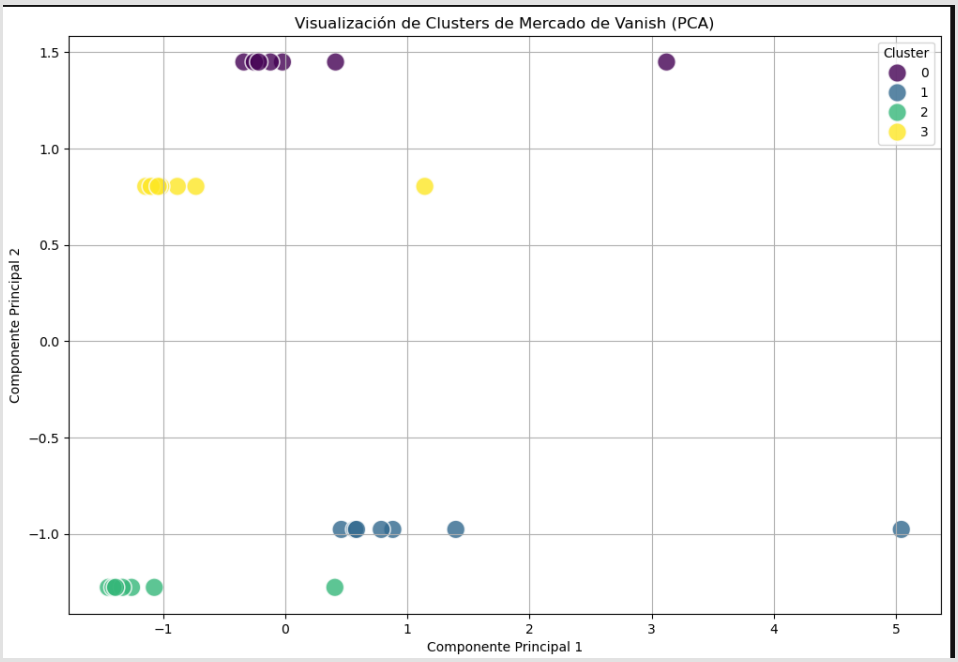
```
--- Análisis Descriptivo de Ventas por Cluster ---
         count           mean            std        min         25%  \
Cluster
0          7.0  100917.715857  113788.402333  35548.220  45598.8555
1          7.0  166867.577714  187370.684604  60905.168  73491.9515
2          7.0    7749.393857    8780.151499   2699.848   3008.2335
3          7.0   33007.945571   36885.860365  12732.436  14306.7250

                50%          75%         max
Cluster
0         55009.604   85728.2380  353212.000
1        100506.316  137820.5800  584036.497
2          3739.197    7333.6860   27122.873
3         20952.123   26614.9245  115527.761
```

- Visualización: Para interpretar los resultados de forma intuitiva, se crearon dos gráficos principales:
    - Un gráfico de barras para comparar las ventas promedio entre los clústers.



    - Un gráfico de dispersión (PCA) para visualizar cómo se separaban los clústers en un plano 2D, facilitando la identificación de patrones.



Análisis de Características por Clúster

El patrón más destacado es que los clústers se definen casi perfectamente por el formato del producto (SEGMENT). Esto indica que el tipo de producto es el factor más influyente en el comportamiento de las ventas, más que la región geográfica.

Clúster 1: Los Líderes del Mercado - Segmento Dominante: **LIQUID & GEL** (Líquidos y Geles).

- **Desempeño en Ventas:** Sobresaliente. Este es el clúster con el promedio de ventas más alto.
- **Presencia Regional:** Fuerte y consistente en todas las regiones analizadas.
- **Patrón Clave:** Representa el negocio principal y más rentable de Vanish. Su alto volumen de ventas lo distingue claramente de los demás formatos.

Clúster 0: Los Pilares Sólidos - Segmento Dominante: **POWDER** (Polvos).

- **Desempeño en Ventas:** Fuerte. Es el segundo clúster con mayor volumen de ventas, consolidándose como un pilar fundamental de la marca.
- **Presencia Regional:** Presente y con buen desempeño en todas las regiones.
- **Patrón Clave:** Al igual que el clúster 1, forma parte del núcleo del negocio. Aunque sus ventas son ligeramente inferiores a las de los líquidos, su rendimiento es muy superior al de los segmentos de nicho.

Clúster 3: El Segmento de Nicho - Segmento Dominante: **BAR** (Barras).

- **Desempeño en Ventas:** Bajo. Sus ventas son considerablemente menores que las de los clústers principales.
- **Presencia Regional:** Distribuido en todas las regiones, pero sin lograr un volumen de ventas significativo en ninguna de ellas.
- **Patrón Clave:** Representa un mercado de nicho. El formato en barra tiene un comportamiento de ventas muy diferente y de menor escala, lo que lo agrupa en un clúster separado y de bajo rendimiento.

Clúster 2: Área de Oportunidad Estratégica- Segmento Dominante: **PRETREAT** (Pretratamiento).

- **Desempeño en Ventas:** Muy Bajo. Este es el clúster con el rendimiento más bajo de todos.
- **Presencia Regional:** Presente en todas las regiones, pero con ventas mínimas.
- **Patrón Clave:** Agrupa a los productos con un caso de uso muy específico que no logran tracción en el mercado. Es el área que requiere la mayor atención estratégica, ya sea para reimpulsar el segmento o para considerar su viabilidad a futuro.

¿Hay segmentos donde la marca Vanish tiene un fuerte desempeño?

Sí, rotundamente. El análisis identifica dos clústers que representan los pilares del negocio de Vanish:

- **Clúster 1 (Líquidos y Geles):** Este es el segmento de más alto desempeño, con las ventas promedio más elevadas (~$166,867). Los productos en formato LIQUID & GEL son los verdaderos motores de venta de la marca en todas las regiones.
- **Clúster 0 (Polvos):** Este clúster, compuesto exclusivamente por el segmento POWDER, es el segundo con mejor rendimiento (~$100,917 en ventas promedio). Junto con los líquidos y geles, forma el núcleo del negocio de Vanish.

Conclusión clave: Los formatos tradicionales y de uso principal (líquidos, geles y polvos) son los que generan la gran mayoría de los ingresos y demuestran la fortaleza de la marca.

¿Qué clústers muestran un rendimiento bajo o áreas de mejora?

el análisis también señala claramente los segmentos con un rendimiento significativamente inferior que requieren atención estratégica:

- **Clúster 2 (Pretratamiento):** Este es el segmento con el rendimiento más bajo de todos (~$7,749 en ventas promedio). Los productos de PRETREAT representan un área de mejora crítica o un punto a revaluar estratégicamente.
- **Clúster 3 (Barras):** Este clúster, que corresponde a los productos en BAR, también muestra un rendimiento bajo (~$33,007 en ventas promedio) en comparación con los segmentos principales.

Conclusión clave: Estos dos clústers representan nichos de bajo rendimiento o áreas donde la estrategia actual no está generando los resultados esperados.

¿Existen patrones en los productos o regiones que influyan en la estrategia comercial?

Sí, el análisis revela un patrón muy claro y potente: el formato del producto (SEGMENT) es el factor más determinante del desempeño en ventas, y este patrón es consistente a través de todas las regiones.

Patrón Identificado:

1. **Patrón de Alto Rendimiento:** Los productos diseñados para el uso principal en el ciclo de lavado (LIQUID & GEL y POWDER) son los que tienen un éxito generalizado y robusto en todas las áreas geográficas. La estrategia para estos segmentos debe ser de protección, innovación y expansión para mantener su liderazgo.
2. **Patrón de Bajo Rendimiento:** Los productos con un caso de uso más específico o de nicho (PRETREAT y BAR) tienen un rendimiento bajo de manera consistente en todas las regiones.
    - **Característica en Común:** La característica que une a los productos de bajo rendimiento es que no son el formato principal de lavado, sino un complemento o un formato alternativo. Su bajo desempeño podría explicarse por varias razones:
        - **Menor Frecuencia de Uso:** Los consumidores pueden no necesitarlos en cada lavado.
        - **Falta de Hábito o Educación:** Puede que los clientes no entiendan completamente los beneficios o cómo incorporar estos productos en su rutina.
        - **Menor Inversión en Marketing:** Es probable que la marca concentre su presupuesto publicitario en los segmentos de mayor volumen (polvos y líquidos).

Implicación Estratégica: Este patrón demuestra que Vanish no puede aplicar una estrategia única para todos sus productos. La compañía necesita un enfoque diferenciado: una estrategia para su negocio principal (polvos y líquidos) y otra muy distinta para sus segmentos de nicho (pretratamiento y barras). Para estos últimos, la pregunta clave es si se debe invertir en educar al consumidor para aumentar su uso o si son productos que se deben optimizar o incluso descontinuar para enfocar recursos en áreas más rentables.

Conclusión

Este análisis demuestra el poder de la ciencia de datos para transformar un gran volumen de información de ventas en una guía estratégica clara y accionable. A través de la consolidación de datos, la ingeniería de características y la aplicación del algoritmo de clustering **K-Means**, se logró segmentar el mercado de Vanish en **cuatro clústers distintos**, cada uno con un perfil de rendimiento y características únicas.

El hallazgo principal es que el desempeño de Vanish no es uniforme; varía significativamente según la combinación de **región geográfica y segmento de producto**. La segmentación reveló:

- **Motores de Crecimiento:** Clústers consolidados y de alto rendimiento que son vitales para el negocio.
- **Segmentos con Potencial:** Mercados con un buen desempeño que, con la inversión adecuada, podrían convertirse en los próximos motores de crecimiento.
- **Nichos y Áreas de Oportunidad:** Segmentos de bajo rendimiento que requieren una evaluación profunda para decidir si se debe invertir, optimizar o descontinuar.

En definitiva, este proyecto va más allá de un simple ejercicio técnico. Proporciona a Vanish una **herramienta estratégica** para dejar atrás un enfoque de "talla única" y adoptar estrategias de marketing, ventas y distribución diferenciadas y personalizadas para cada tipo de mercado. La capacidad de identificar con precisión dónde enfocar los recursos es clave para optimizar la inversión, maximizar la rentabilidad y asegurar una ventaja competitiva sostenible.