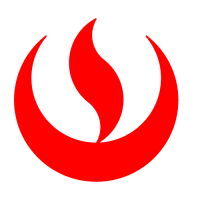
**UNIVERSIDAD PERUANA DE CIENCIAS APLICADAS**

**FACULTAD DE INGENIERÍA**

**CARRERA PROFESIONAL**

**DE CIENCIAS DE LA COMPUTACIÓN**



**Asignatura:**

**CC57- APLICACIONES DE DATA SCIENCE**

**Sección: CC73**

**Trabajo Final**

**Sistema de recomendaciones para un E- commerce**

**Autores**

Franck Goñas Lopez – u201421134

**Profesor**

Carlos Fernando Montoya Cubas

Lima, abril de 2024

**CONTENIDO**

[1. Descripción del caso de uso 3](#_30j0zll)

[2. Descripción del conjunto de datos (dataset) 4](#_1fob9te)

[3. Análisis exploratorio de los datos (EDA) 4](#_3znysh7)

[4. Propuesta de Modelización 10](#_2et92p0)

[5. REFERENCIAS BIBLIOGRÁFICAS 11](#_1t3h5sf)

[6. ANEXOS 12](#_l19bdzymncpe)

# Descripción del caso de uso

En el contexto del comercio electrónico y las tiendas minoristas, uno de los retos más significativos es mantener la relevancia y personalización en la interacción con cada cliente. Con una vasta cantidad de productos disponibles y una diversidad de preferencias de los consumidores, es esencial para las empresas entender y anticipar las necesidades de sus clientes para mejorar la experiencia de compra y optimizar las ventas. El problema fundamental es cómo maximizar la eficacia del proceso de venta cruzada y la venta sugerida de productos sin abrumar al cliente con opciones irrelevantes. Un sistema efectivo debe ser capaz de predecir las preferencias futuras y presentar recomendaciones de producto.

Este escenario presenta una oportunidad significativa para desarrollar un sistema de recomendación que no solo aumente la satisfacción y fidelidad del cliente, sino que también impulse las ventas incrementales a través de recomendaciones personalizadas y contextualmente apropiadas.

# Descripción del conjunto de datos (dataset)

El dataset usado en el presente trabajo se extrajo de la plataforma web Kaggle, siendo más específico de <https://www.kaggle.com/datasets/carrie1/ecommerce-data/data>

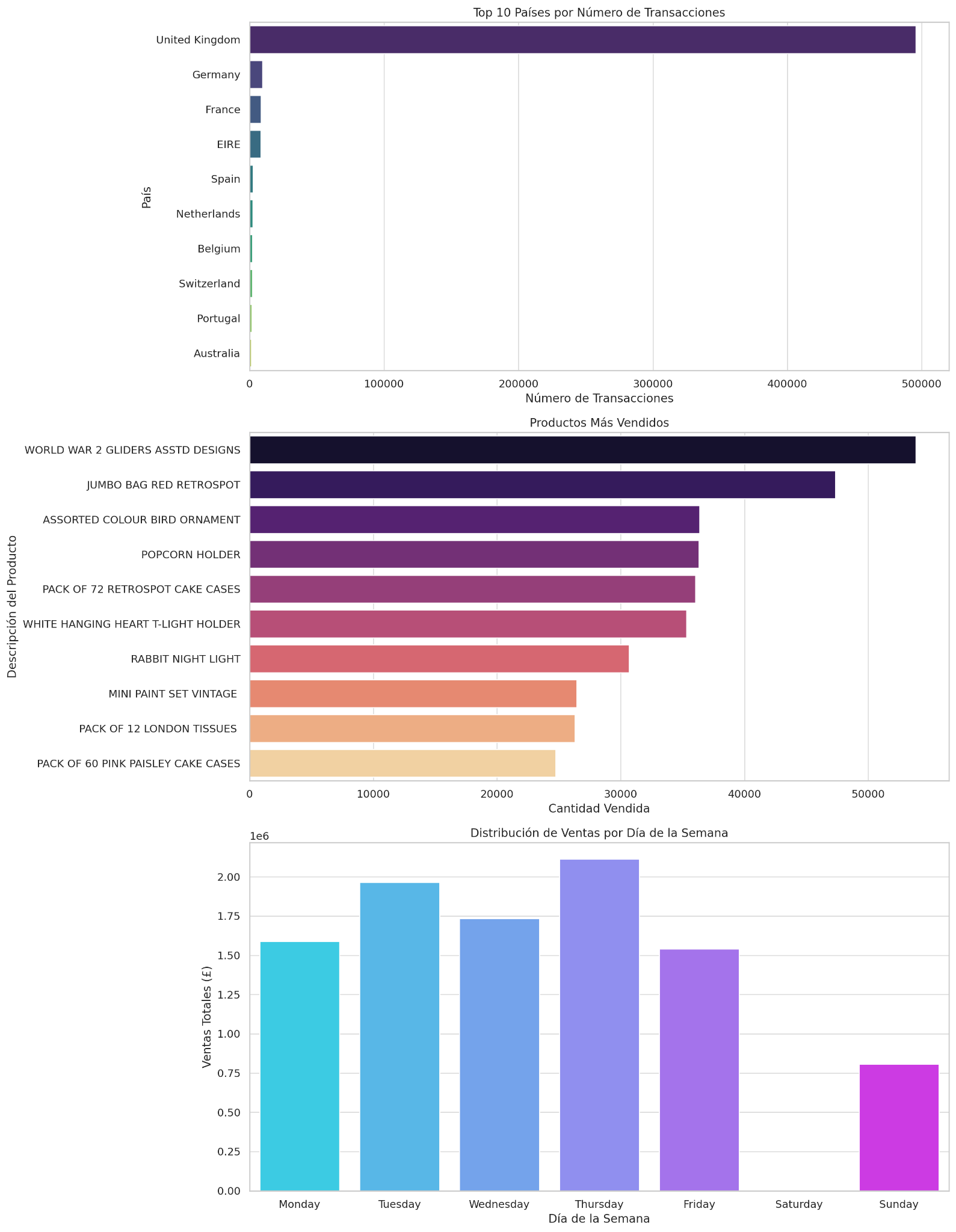
El conjunto de datos contiene 541,909 registros distribuidos en 8 columnas.

| Variable | Tipo de Dato | Descripción |
| --- | --- | --- |
| InvoiceNo | int | Número de factura o transacción. Identificador único para cada transacción. |
| StockCode | int | Código del producto. Código único asignado a cada producto diferente en el inventario. |
| Description | text | Descripción textual del producto. Puede contener detalles como color, material, etc. |
| Quantity | int | Cantidad del producto en la transacción. Los números negativos podrían indicar devoluciones. |
| InvoiceDate | date | Fecha y hora de la factura. Formato de fecha y hora cuando se realizó la transacción. |
| UnitPrice | float | Precio por unidad del producto en GBP. |
| CustomerID | int | Identificador único del cliente. |
| Country | text | País del cliente. Indica la ubicación geográfica desde donde se realizó la transacción. |

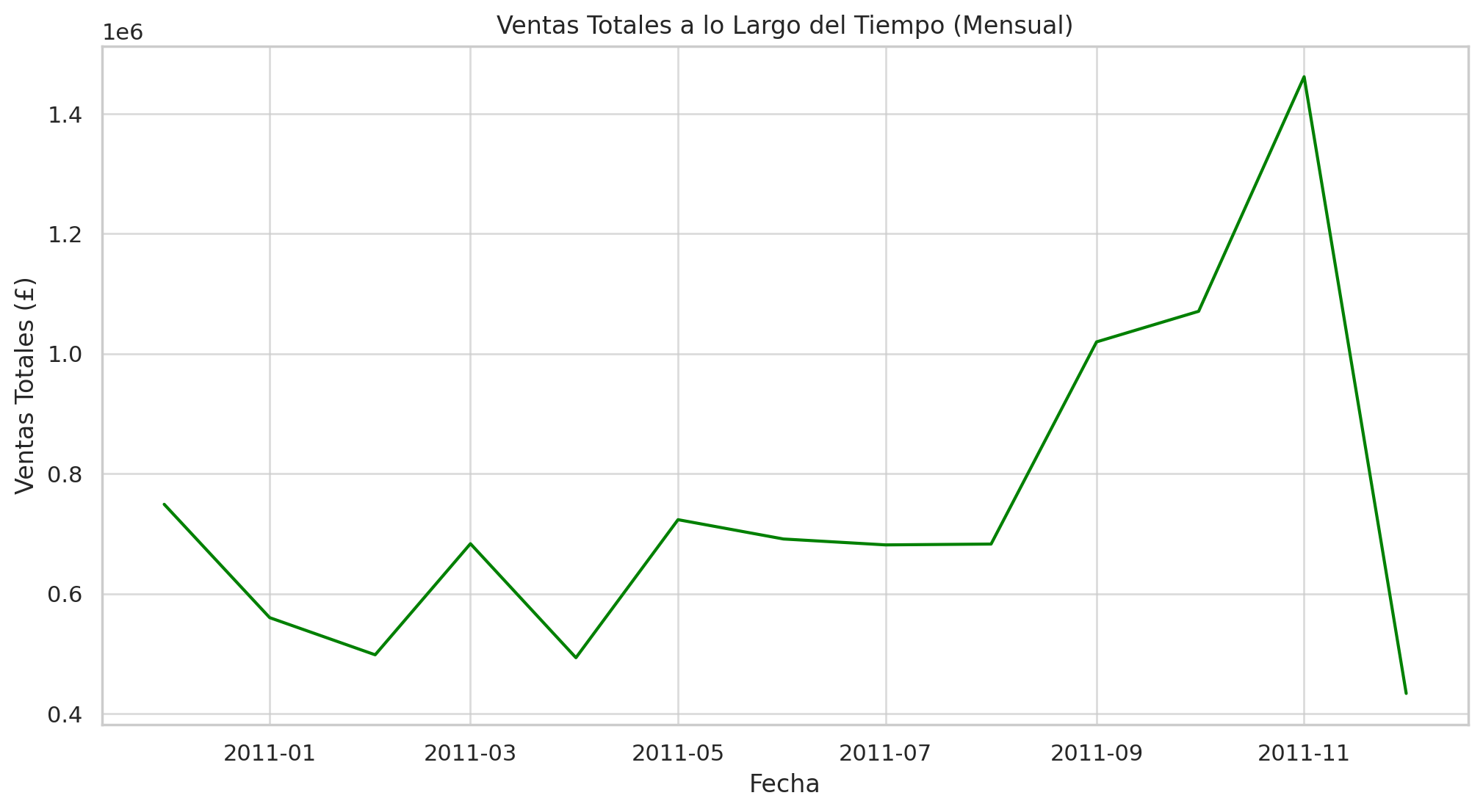
# Análisis exploratorio de los datos (EDA)

Se puede acceder al código en el Anexo 1.

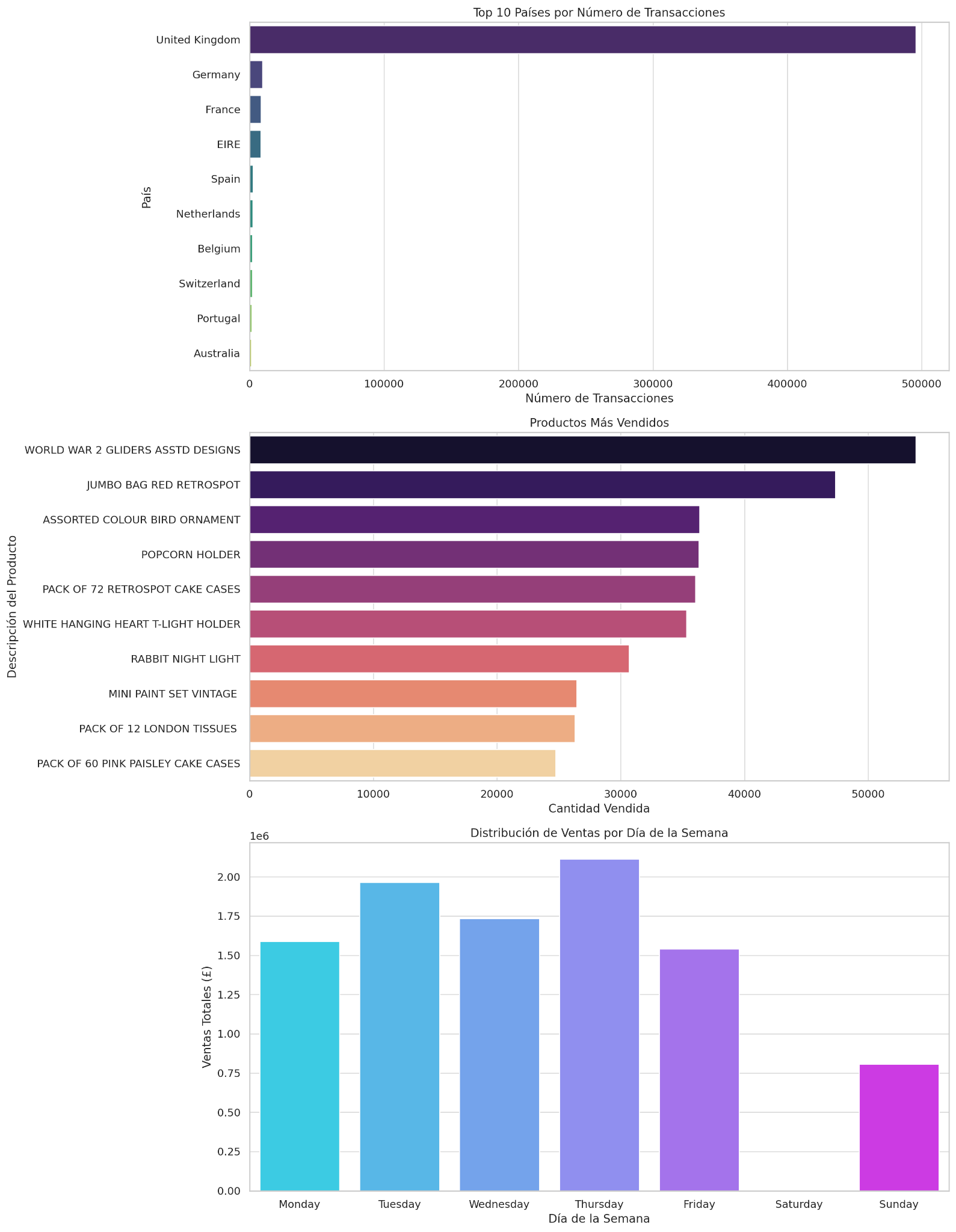
Para el análisis de datos se comenzó explorando la distribución de las ventas en la variable país.



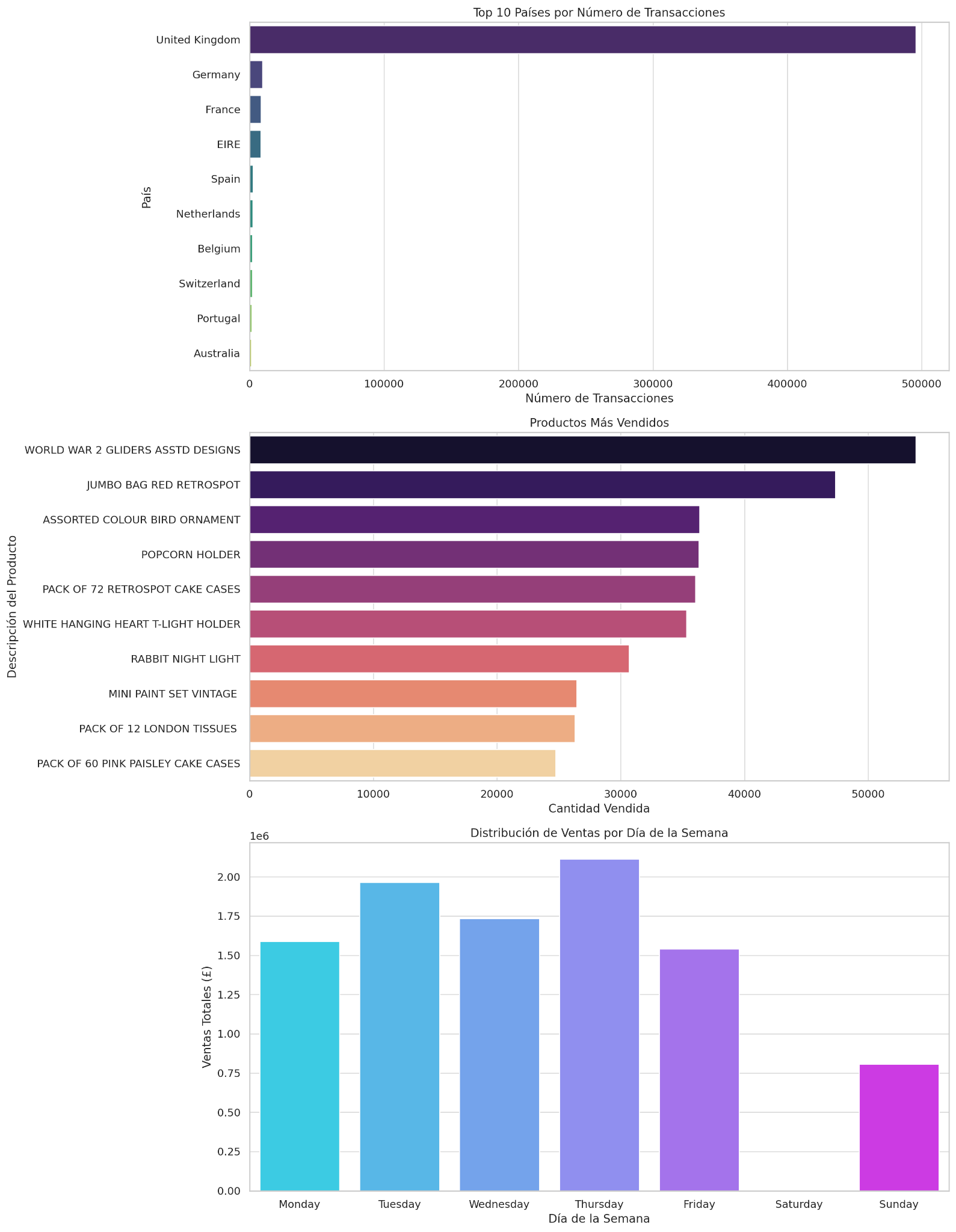
Podemos ver que la gran mayoría de compras se hicieron en United Kingdom.



Esta gráfica ayuda a observar las fluctuaciones mensuales en las ventas totales.



En este gráfico podemos notar los 10 productos más vendidos por la empresa de e-commerce.



Esta gráfica ilustra cuáles días de la semana generan más ventas, ayudando a planificar promociones y operaciones en función de la demanda.

# Propuesta de Modelización

La modelización del sistema de recomendación comenzó con la selección de algoritmos adecuados para ser entrenados a partir de los datos preparados. En este caso, se optó por utilizar el algoritmo de Descomposición de Valores Singulares (SVD) para la recomendación colaborativa y KMeans para el agrupamiento de usuarios y productos, complementado con técnicas de procesamiento de datos y normalización.

Preparación y Limpieza de Datos:

Los datos se cargaron desde un archivo CSV y se limpiaron eliminando valores nulos y convirtiendo los tipos de datos según corresponda.

Se creó una columna de frecuencia para representar el número de veces que un usuario ha comprado un producto específico.

Creación de la Matriz de Usuario-Producto:

Se creó una matriz de usuario-producto basada en una función con la frecuencia de compra utilizando la biblioteca surprise. La funcion utilizada mantiene los valores 0 en 0, los 1 en 1 y los valores mayores a 1 los pone en un rango hasta el 5 logaritmicamente.

El modelo SVD se entrenó y evaluó utilizando la métrica RMSE para medir el rendimiento.

Agrupamiento de Usuarios y Productos:

Se utilizó KMeans de cuML para el agrupamiento de usuarios y productos basado en las frecuencias de compra, probó con diferentes cantidades de grupos y se optó por 1000 para los grupos de usuario y 100 para los grupos de productos.

Los clústeres generados se asignaron a los usuarios y productos en el DataFrame original.

Generación de Recomendaciones:

Se desarrollaron funciones para obtener recomendaciones basadas en los clústeres de usuarios y productos, productos mas vendidos, productos últimamente más vendidos, recomendaciones basadas en el modelo svd y se implementaron recomendaciones híbridas que combinan ambos enfoques(cluster y svd).

véase con más detalle en el [Anexo1](#_pcl6xz57smgs)

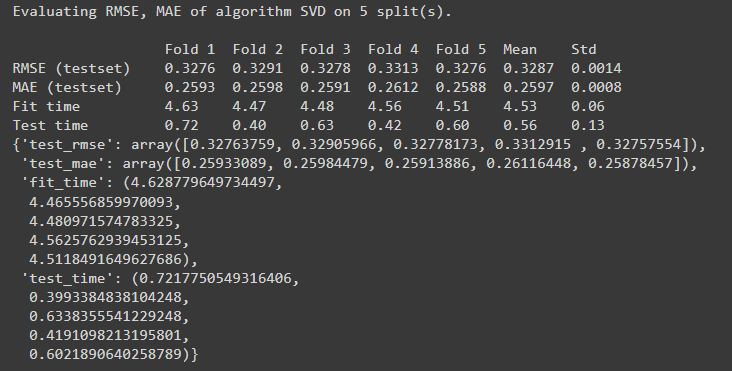
# Publicación de los resultados.

Los resultados de los experimentos realizados con los modelos creados se comunicaron utilizando las métricas RMSE y MAE para evaluar el rendimiento del modelo SVD. A continuación, se presentan los resultados obtenidos:

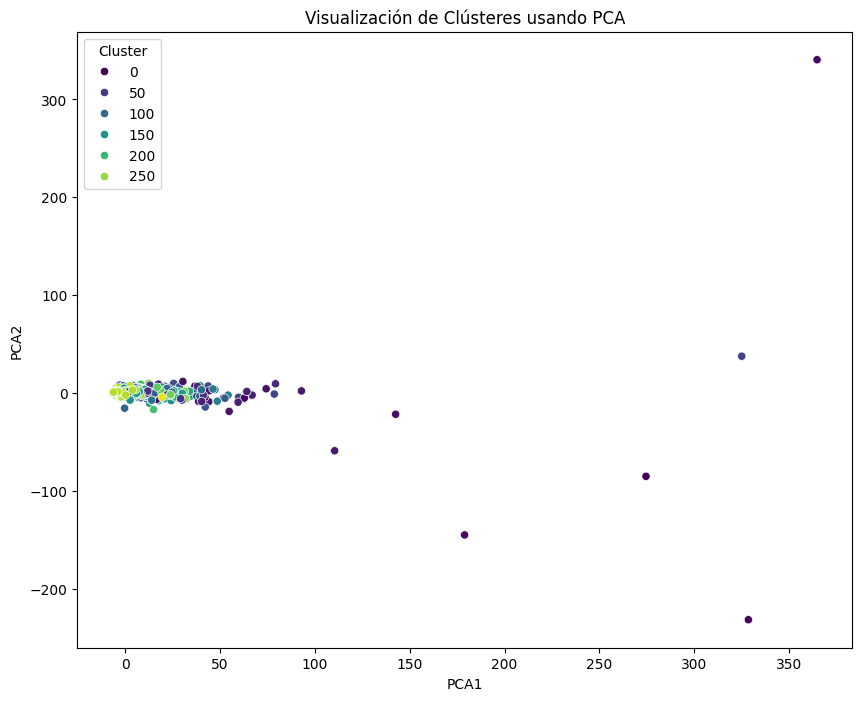
RMSE del modelo SVD final:



Validación cruzada:



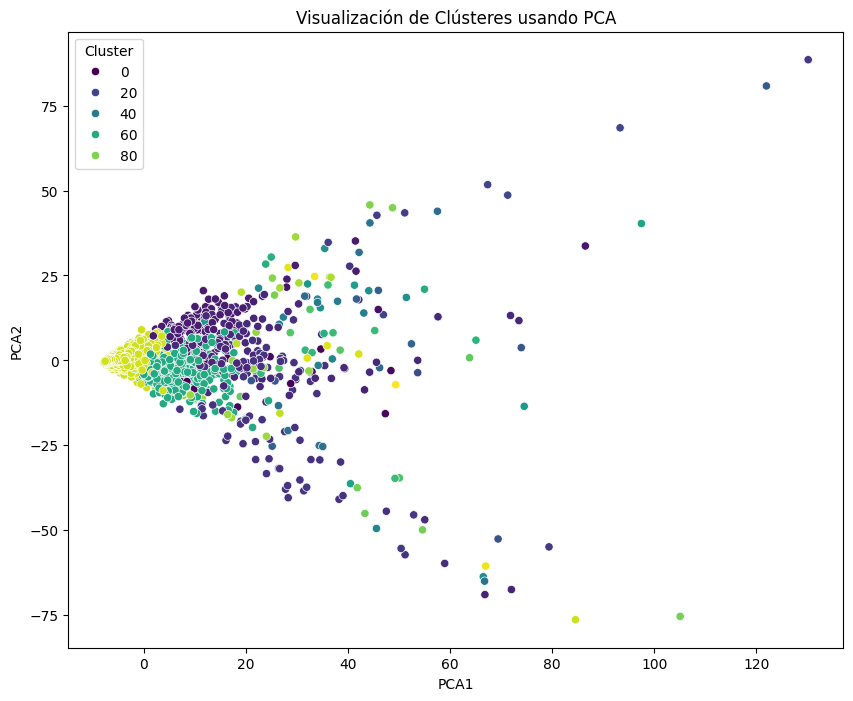
Por otro lado, el modelo de clusterización para clientes con mayor a 5 compras de productos(el rendimiento de la clusterizacion mejora mientras mayor sea la cantidad mínima de ventas para entrenar el modelo):





Indica una calidad de clustering moderadamente baja

Clusterización de productos:





Indica una calidad de clustering moderada

**Recomendaciones**

Posterior a esto se usan las clusterizaciones y el modelo svd para hacer recomendaciones(se brindan los codigo de producto los cuales serían los que se brindaría al api de la web)

Recomendaciones de productos más vendidos y mas vendidos recientemente(sobretodo para usuarios que no tienen 10 compras aún)



Ejemplo para el usuario con id: **17850**

Recomendaciones basadas en clusters de usuarios(considerando lo que más compran los usuarios parecidos):

Recomendaciones basadas en clusters de ítems o productos:



Recomendaciones según el modelo svd:



Recomendaciones hibridas(usan los grupos de usuarios, productos y el modelo svd)



# Análisis de Explicabilidad.

Para el análisis de su uso factores latentes en una predicción específica.

Predicción Específica

Usuario: 17850

Ítem: 23084

Predicción: 0.14047709059168723

Factores Latentes

Factores Latentes del Usuario 17850:

[ 0.0652586, 0.06324102, 0.15576508, -0.08209931, 0.04968939, ...]

Factores Latentes del Ítem 23084:

[ 0.13605953, 0.06421362, -0.07442971, -0.07988153, 0.03758756, ...]

Contribución de Cada Factor a la Predicción

Cada factor latente representa una dimensión de características subyacentes tanto para los usuarios como para los productos. La predicción es el resultado del producto de los factores latentes del usuario y del ítem.

Los factores con las mayores contribuciones absolutas son los más relevantes para la predicción. A continuación se muestran los factores más significativos:

Índice del Factor Contribución

17 0.0344838207

21 0.00467512704

40 0.0177046795

43 0.0109324880

57 0.0269257366

Estos factores tienen las mayores contribuciones absolutas y, por lo tanto, son los más influyentes en la predicción para este usuario y producto específicos.

# Conclusiones

En este proyecto, se utilizaron técnicas de recomendación colaborativa y de contenido, así como agrupamiento basado en frecuencia de compras. Se entrenó un modelo SVD para recomendaciones colaborativas y se utilizó KMeans para agrupar usuarios y productos. Los resultados obtenidos muestran un buen rendimiento del modelo SVD con un RMSE aceptable, y las recomendaciones generadas son coherentes con los patrones de compra observados.

Para futuros trabajos, se puede considerar adquirir más datos para la implementación de modelos más avanzados como redes neuronales recurrentes (RNN) y técnicas de aprendizaje profundo para mejorar aún más las recomendaciones. Además, se pueden explorar técnicas de explicabilidad más sofisticadas para entender mejor las decisiones del modelo y aumentar la confianza de los usuarios en las recomendaciones proporcionadas.

# Bibliografía

Bobadilla, J., Ortega, F., Hernando, A., & Bernal, J. (2012). A collaborative filtering approach to mitigate the new user cold start problem. *Knowledge-Based Systems, 26*, 225-238. <https://doi.org/10.1016/j.knosys.2011.07.021>

FARZAD NEKOUEI(2022). *Customer Segmentation & Recommendation System*. Kaggle. Recuperado el 20 de junio de 2024, de<https://www.kaggle.com/datasets/carrie1/ecommerce-data/data>

Carrie (2015). *E-commerce data*. Kaggle. Recuperado el 20 de junio de 2024, de<https://www.kaggle.com/datasets/carrie1/ecommerce-data/data>

Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. In *Recommender Systems Handbook* (pp. 1-35). Springer. <https://doi.org/10.1007/978-0-387-85820-3_1>

# ANEXOS

Anexo1: Repositorio github

<https://github.com/franckhbz/-CC219-TP-TF-2024-1>