

Grado en

**Business Data Analytics**

Informe final

Reto: 4 Eroski

Curso: 1º 2024-2025

Equipo: Azul Claro

URL del repositorio en GitHub:

Autores:

* Ander Sainz de la Maza
* Imanol Egiguren
* Erik Arbaiza
* Naroa Laria
* Ander Oquina
* Olatz Urkidi

24-25\_R4\_AzulClaro

# Índice

Contenido

[Índice 2](#_Toc195007235)

[1 Introducción 3](#_Toc195007236)

[2 Identificar la problemática 4](#_Toc195007237)

[2.1 Sobre el cliente 4](#_Toc195007238)

[2.2 Sobre la problemática 4](#_Toc195007239)

[2.3 Objetivos 5](#_Toc195007240)

[3 Recogida y procesado de los datos 6](#_Toc195007241)

[3.1 Fuentes de datos utilizadas 6](#_Toc195007242)

[3.2 Procesamiento de los datos 6](#_Toc195007243)

[4 Análisis y modelización de los datos 7](#_Toc195007244)

[4.1 Análisis de los datos 7](#_Toc195007245)

[4.2 Clustering de Clientes 7](#_Toc195007246)

[4.3 Modelización 7](#_Toc195007247)

[4.4 Conclusiones 7](#_Toc195007248)

[*5* Transformar los negocios 8](#_Toc195007249)

[6 Implicaciones legales y éticas 9](#_Toc195007250)

[Bibliografía 10](#_Toc195007251)

## Introducción

Este informe se estructura diferentes apartados orientados al análisis de diferentes empresas. El primer apartado describe al cliente, en este caso a Eroski, y su rol estratégico en la competencia. Aquí se habla de la historia y de la empresa como tal, al igual que se presenta la problemática obtenida y los objetivos específicos a los que se enfrenta.

Posteriormente se aborda la recopilación y almacenamiento de datos, usando tanto fuentes internas como externas. Específicamente, se trabaja con…. Este bloque detalla el proceso de limpieza y transformación de datos, esencial para garantizar la coherencia en el análisis.

Después, se explora la fase de análisis y modelado de los datos. Se inicia con un análisis descriptivo ….

Más adelante, se verán las recomendaciones que se sugieren hacer a la empresa para mejorar su negocio y hacer frente a posibles obstáculos de un futuro. Esto ayudará en la optimización de sus recursos para hacer un mejor uso.

Finalmente, el informe incluye una bibliografía en la última página, donde se especifican todas las fuentes de información empleadas. Esta sección no solo respalda la credibilidad del informe, sino que también ofrece transparencia sobre los datos y métodos utilizados.

## Identificar la problemática

### Sobre el cliente

Eroski es una empresa cooperativa española conocida por una de las principales cadenas de supermercados en España. Se fundó en 1969 en el País Vasco como parte del grupo cooperativo Mondragón y ha crecido hasta convertirse en una de las empresas líderes en el sector de la distribución alimentaria.

Su modelo cooperativo se basa en que los trabajadores y socios consumidores pueden participar en la gestión y toma de decisiones de la empresa. Esto permite que la compañía mantenga un fuerte compromiso con sus empleados y clientes, favoreciendo un modelo de negocio más participativo y socialmente responsable.

Eroski tiene una fuerte implantación en el norte del país, especialmente en regiones como el País Vasco, Navarra y Galicia. Sin embargo, también opera en otras comunidades autónomas a través de diferentes formatos de tiendas. Entre estos formatos se incluyen hipermercados, tiendas de conveniencia y franquicias como Caprabo. Además, también cuenta con estaciones de servicio y tiendas especializadas en productos de viaje y ocio.

Aun así, tienen un gran compromiso con la sostenibilidad y la responsabilidad social. Ha implementado distintas iniciativas para reducir el impacto ambiental, como la promoción de productos ecológicos, el fomento de productos locales y la reducción del uso de plásticos en sus establecimientos. Colabora con organizaciones y programas solidarios para apoyar a colectivos vulnerables. No solo se dedica a la distribución alimentaria, sino que, también apuesta por un modelo de negocio cooperativo, sostenible y socialmente comprometido.

### Sobre la problemática

En la actualidad, el comercio electrónico enfrenta el desafío de proporcionar experiencias de compra personalizadas y eficientes para sus clientes. Con el crecimiento exponencial de los datos transaccionales, surge la necesidad de desarrollar sistemas de recomendación que permitan optimizar la interacción entre los usuarios y los productos disponibles.

Eroski, como empresa de distribución, busca mejorar la experiencia de compra de sus clientes a través de un sistema de recomendaciones basado en sus transacciones. Para lograrlo, es necesario analizar los datos históricos de compras con el objetivo de identificar patrones de consumo y generar recomendaciones precisas y relevantes.

El desarrollo de estos sistemas de recomendación implica un proceso complejo de análisis y procesamiento de datos, que incluye la identificación de patrones de compra, la reducción de dimensionalidad para optimizar el rendimiento de los modelos y la creación de una matriz de datos representativa de las relaciones entre clientes y productos.

En este contexto, el principal desafío radica en la capacidad de extraer información valiosa a partir de grandes volúmenes de datos transaccionales y transformarla en recomendaciones útiles y personalizadas. Esto no solo mejorará la experiencia del cliente, sino que también contribuirá al incremento de las ventas y la fidelización de los usuarios.

### Objetivos

Este reto presenta cuatro escenarios específicos en los que se requieren distintos tipos de recomendaciones:

1. Artículo promocionado: Se busca identificar los 10 clientes más adecuados para recibir la recomendación de un producto en particular, sin restricciones sobre compras previas.
2. Otros como tú han comprado: Se pretende sugerir a cada uno de los 10 clientes un ítem adicional que no esté en su carrito, basándose en patrones de compra similares de otros usuarios.
3. Oferta para ti: Se recomienda a cada cliente un artículo dentro de un conjunto de 20 productos en oferta, seleccionando el más adecuado según su historial de compras.
4. Quizás te hayas olvidado: Se busca identificar productos que los clientes podrían haber olvidado agregar a su cesta, con base en su última compra.

Por lo tanto, este proyecto busca abordar la problemática de la personalización en el comercio electrónico, utilizando estrategias avanzadas de análisis de datos para optimizar las recomendaciones de compra y generar un impacto positivo en la relación entre Eroski y sus clientes.

## Recogida y procesado de los datos

### Fuentes de datos utilizadas

*Breve descripción de las bases de datos proporcionadas*

### Procesamiento de los datos

*Explicación de la selección de criterios para la reducción de dimensionalidad*

*Descripción del procesamiento necesario para crear la matriz de datos*

## Análisis y modelización de los datos

### Análisis de los datos

*Análisis estadístico-descriptivo de las variables utilizadas*

*Gráficos descriptivos de las variables relevantes*

*Conclusiones generales y particulares del análisis*

(Intro)

Gráfico, Histograma

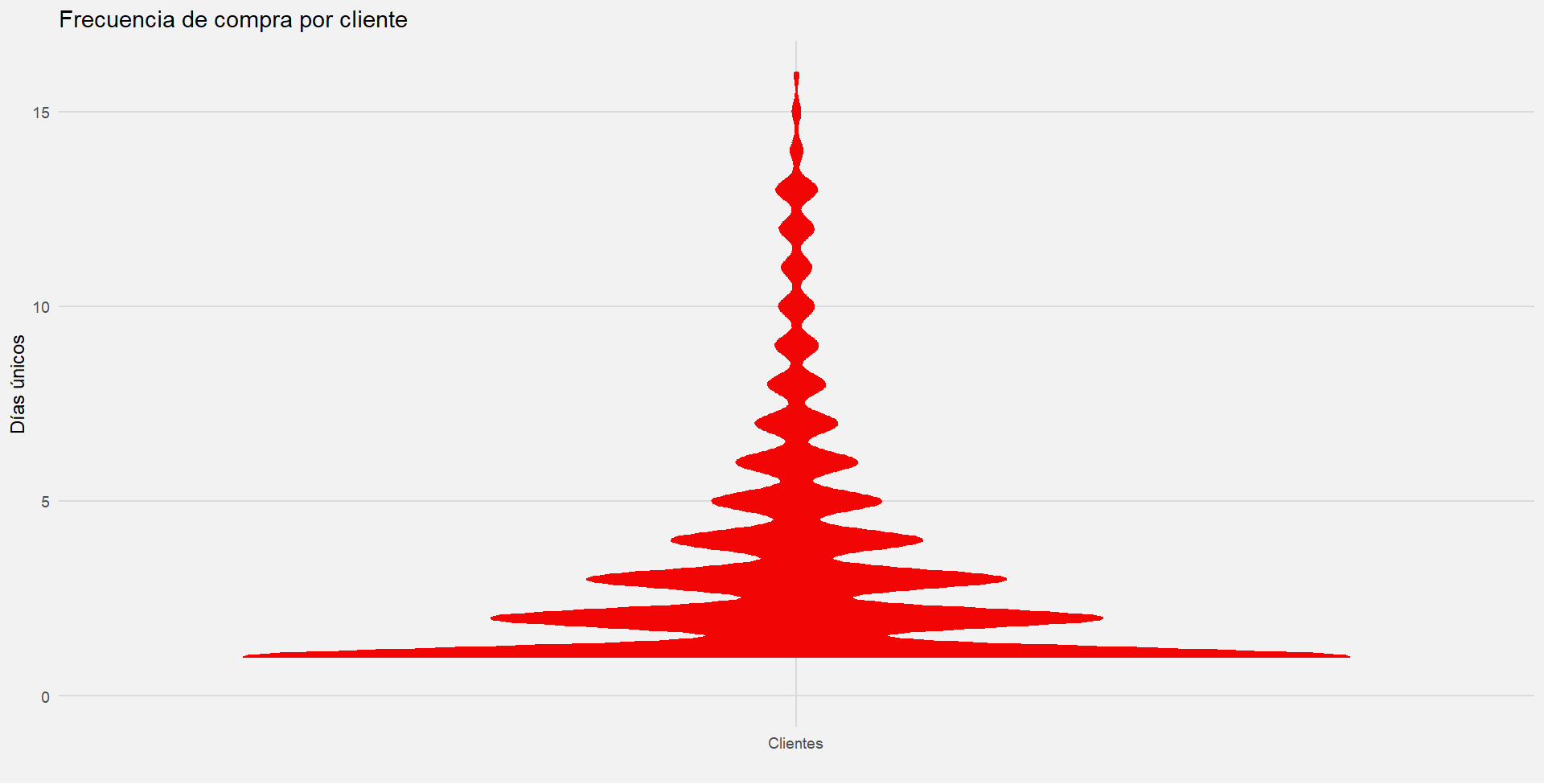
El contenido generado por IA puede ser incorrecto.

El gráfico muestra la distribución del número de productos distintos adquiridos por cliente en el periodo analizado. La forma de la distribución es notablemente asimétrica reflejando una clara concentración que nos indica que la mayoría de los clientes escoge una cantidad limitada de productos distintos, mientras que sólo una poca proporción muestran comportamientos de compra más diversificados.

Existe un rango que refleja una mayor frecuencia de clientes. Este se encuentra entre los 40 y 50 productos diferentes, donde se llega a superar la cifra de 10,000 clientes. Esto permite concluir que el cliente promedio suele mantener un nivel moderado de variedad en sus compras. A medida que aumenta el número de productos obtenidos, la cantidad de clientes decrece de forma sostenida, mostrando que los comportamientos de compra diversos son inusuales.

En los extremos de la distribución se observan casos atípicos que indican que algunos clientes han adquirido más de 200 productos diferentes, alcanzando incluso valores por encima de los 500. Aunque hay muy pocos casos, puede ser relevante ya que podrían identificar a aquellos a clientes que pertenecen a empresas o consumidores intensivos.

Este tipo de análisis es fundamental para la segmentación estratégica de la base de clientes debido a que los clientes que muestran un alto grado de diversidad en sus compras representan oportunidades para programas de fidelización o beneficios exclusivos, dado su potencial valor a largo plazo para la compañía. Además, esta distribución aporta una visión clara del comportamiento unido a  los clientes y permite diseñar estrategias comerciales diferenciadas y efectivas en función del nivel de diversidad de consumo.

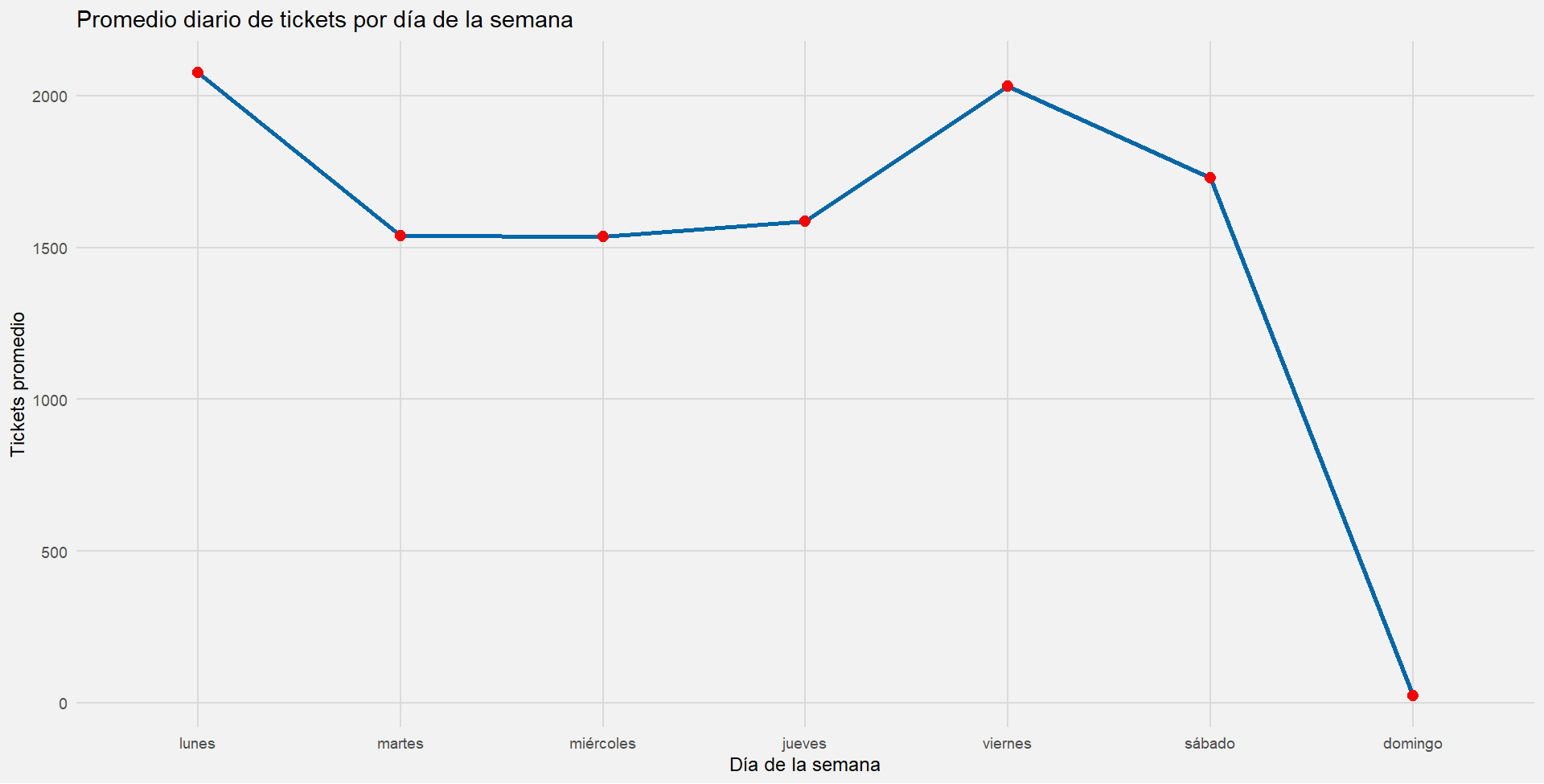


El gráfico presentado muestra la frecuencia de compra por cliente, a través de los días en los que cada cliente realizó alguna transacción durante el periodo de análisis. Visualmente, la distribución refleja una forma simétrica, con un claro pico central que indica una mayor concentración de clientes en torno a una frecuencia de compra específica.

Se observa que la mayoría de los clientes realizaron compras en aproximadamente cinco días únicos, lo que representa el valor más frecuente dentro del conjunto de datos. A partir de ahí, la cantidad de clientes disminuye tanto hacia frecuencias menores como hacia frecuencias mayores, formando una curva indicando una distribución normal. Este patrón indica que la frecuencia de compra tiende a estar concentrada en torno a un rango medio, con pocos clientes realizando compras en muy pocos o en muchos días distintos.

El hecho de que la mayoría de los clientes tengan una frecuencia de compra baja puede sugerir un comportamiento de compra ocasional. Sin embargo, existe un grupo más reducido de clientes que realizó compras en un mayor número de días, reflejando un mayor grado de fidelidad, por ejemplo.

Este gráfico permite identificar segmentos de clientes que podrían beneficiarse de estrategias de marketing diferenciadas. Los que cuentan con una baja frecuencia de compra pueden ser incentivados mediante campañas de activación o promociones limitadas. Por otro lado, los clientes con una alta frecuencia representan una oportunidad para fortalecer su lealtad a través de beneficios exclusivos.

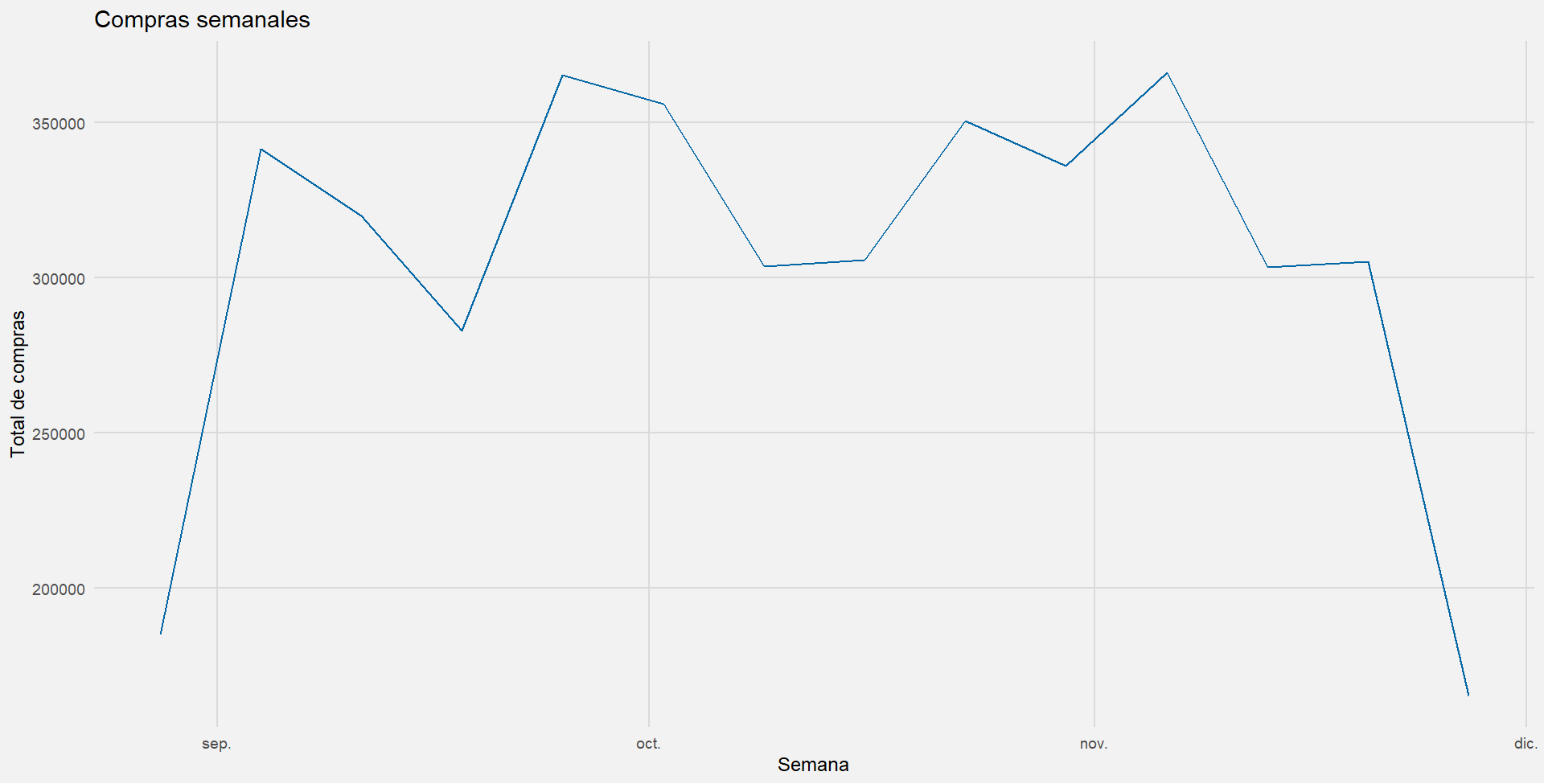


El gráfico muestra el promedio diario de tickets generados por día de la semana, permitiendo identificar patrones de comportamiento en la actividad de compra a lo largo del ciclo semanal. Se observa una variación significativa en la cantidad promedio de tickets según el día, por lo que ciertos días concentran una mayor actividad comercial que otros.

Los lunes y viernes presentan los niveles más altos de tickets, superando los 2,000 tickets promedio diarios. Esta situación podría estar asociada a patrones típicos de consumo. El sábado también cuenta con un nivel relativamente elevado de actividad, aunque algo menor que el de los días mencionados, lo cual podría vincularse al tiempo libre disponible de los consumidores.

Sin embargo, el martes, el miércoles y el jueves muestran una estabilidad en la cantidad promedio de tickets, con valores más bajos, que giran en torno a los 1,550 tickets diarios. Por último, el domingo destaca por ser el día con la menor actividad comercial del ciclo semanal, pudiendo producirse por descanso dominical.

Conocer los días de mayor afluencia permite optimizar la asignación de recursos, ajustar la disponibilidad de personal y diseñar campañas promocionales específicas para aumentar la demanda.



La figura anterior muestra el comportamiento de las compras semanales, con base en los datos visualizados a través de un gráfico de línea. Este análisis permite identificar patrones de comportamiento de los consumidores, así como inferencias sobre sus hábitos de compra a lo largo del tiempo.  La variable principal analizada es el total de compras por semana, y el período observado abarca desde el 1 de septiembre de 2022 hasta el 30 de noviembre de 2022, cubriendo un total de aproximadamente tres meses de actividad comercial. La distribución de las semanas permite identificar la estacionalidad del comportamiento de compra, clave para la interpretación de los patrones de consumo.

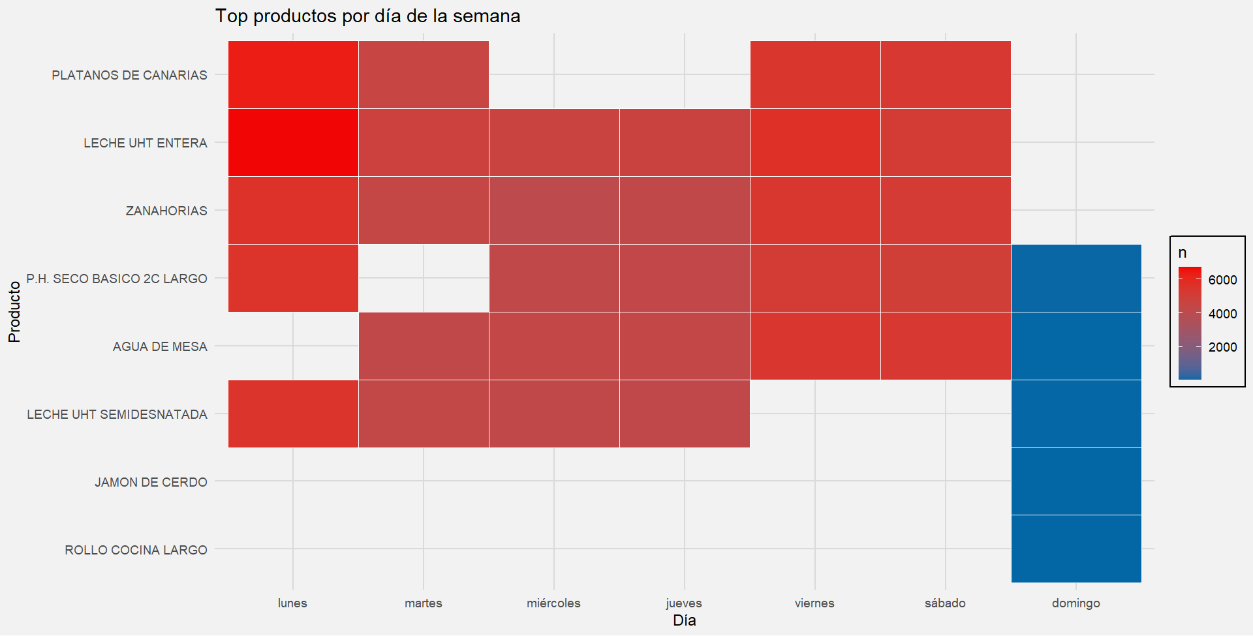
El valor de la variable de compras semanales muestra una media de 306,170 compras, con una mediana ligeramente superior (312,848), lo que sugiere una leve asimetría negativa en la distribución. El rango de compras semanales varía significativamente, desde un mínimo de 165,299 hasta un máximo de 366,022, lo cual indica una variabilidad considerable en la actividad de compra.

Respecto a las tendencias, se aprecia un aumento inicial pronunciado en las primeras semanas de septiembre, alcanzando uno de los valores más altos de todo el período. Posteriormente, se observan fluctuaciones moderadas durante octubre, con semanas de mayor estabilidad en el volumen de compras. Durante noviembre, las compras se mantienen en niveles altos, con un pico a mediados del mes, antes de descender bruscamente en la última semana registrada.

El comportamiento de las compras semanales presenta una tendencia general fluctuante, sin una estacionalidad claramente marcada, pero con ciertos picos destacados. La variabilidad en el número de compras semanales es alta, con diferencias que superan las 200,000 compras entre la semana con menor y mayor actividad. Las semanas de mayor actividad superan el umbral de 360,000 compras, lo cual representa una capacidad operativa elevada.

El análisis estadístico y gráfico de los hábitos de compra en Eroski muestra una actividad comercial intensa y variable a lo largo de las semanas analizadas. Con una media de más de 306.000 compras semanales, los clientes presentan un comportamiento de consumo constante, aunque con picos destacados en septiembre y noviembre. Estos incrementos pueden vincularse a momentos clave del calendario comercial, como la vuelta a la rutina o campañas promocionales como el Black Friday.

Asimismo, la caída observada al final del periodo sugiere la posible influencia de factores como semanas incompletas, festivos o menor actividad promocional. En general, los patrones indican que los consumidores de Eroski responden de forma clara a estímulos temporales y campañas, lo cual representa una oportunidad para optimizar las acciones de marketing y fidelización en momentos estratégicos del año.



El gráfico presenta un heatmap que identifica los productos más populares en Eroski para cada día de la semana. Se observa una clara preferencia de los consumidores por productos de primera necesidad, especialmente lácteos y frutas, a lo largo de toda la semana. Además, también se refleja la intensidad de compra por día mediante el color (cuanto más rojo mayor volumen y cuanto más azul menos compras).

Los productos con mayor frecuencia de compra son:

* Leche UHT entera y semidesnatada: con alta demanda de lunes a sábado.
* Plátanos de Canarias: frecuentes al inicio y final de la semana, posiblemente relacionados con reposición de frutas frescas.
* Zanahorias: muy consistentes en todos los días hábiles.
* P.H. seco básico 2C largo y agua de mesa: aparecen de forma relevante en varios días, con el agua destacando especialmente los fines de semana.

El domingo muestra una menor variedad de productos, donde destacan artículos puntuales como agua de mesa, jamón de cerdo y rollo de cocina largo, lo que sugiere un patrón de compras más orientado a la reposición rápida o emergente. Aunque, cabe destacar que solo se abren las tiendas los domingos en localidades muy concretas, por lo que tampoco se puede extraer una conclusión definitiva sobre los hábitos dominicales a nivel general. A eso se debe también el menor volumen de compras.

El análisis de los productos más comprados en Eroski a lo largo de la semana revela patrones consistentes de consumo, centrados principalmente en productos de primera necesidad, como ya se ha comentado anteriormente. Además, todos ellos están presentes con regularidad en el top de ventas de lunes a sábado.

Asimismo, se ha visto que el lunes registra el mayor volumen de compras en general, con cinco productos superando las 5.000 unidades vendidas. Esto sugiere que los consumidores planifican la semana iniciando con la reposición de productos básicos. Mientras que de martes a viernes se mantiene una compra consistente de productos como leche, agua de mesa, zanahorias y pan de molde (P.H. seco básico 2C largo), lo que refleja hábitos estables de consumo durante la semana laboral. A su vez, el sábado conserva un volumen elevado, especialmente en agua de mesa y fruta fresca, lo que puede asociarse a compras para el fin de semana o antes del cierre dominical. Finalmente, como ya se ha mencionado, los domingos muestran una baja actividad y compras de productos específicos como el jamón de cerdo, el rollo de cocina largo y el agua de mesa. Esta baja actividad se explica tanto por la limitación de aperturas en determinadas localidades como por un patrón de compra centrado en necesidades urgentes o imprevistas.

Gráfico, Histograma

El contenido generado por IA puede ser incorrecto.

El gráfico presentado muestra la distribución de los días transcurridos entre compras sucesivas realizadas por los clientes de Eroski. Este análisis permite identificar patrones de comportamiento en cuanto a la frecuencia con la que los consumidores acuden a realizar sus compras, proporcionando información clave para la planificación comercial y logística.

Se observa un marcado pico de frecuencia en el día 7, lo cual indica que una parte significativa de los clientes realiza sus compras de forma semanal. Este comportamiento sugiere una rutina establecida, probablemente vinculada a la organización semanal del hogar, como la compra de productos frescos y de consumo diario al inicio o final de la semana.

Además, se identifica otro pico relevante en torno al día 14, lo que sugiere la existencia de un segundo grupo de consumidores con un patrón de compra quincenal. Este grupo podría corresponder a hogares que planifican con mayor anticipación y que realizan compras más voluminosas, reduciendo así la frecuencia de sus visitas al establecimiento.

En la parte inicial del gráfico también se evidencia una proporción considerable de compras realizadas con una frecuencia menor a una semana, es decir, entre 1 y 6 días. Este comportamiento responde probablemente a compras de reposición o de productos perecederos, y puede reflejar una clientela que realiza compras más impulsivas o espontáneas.

Los resultados evidencian que el hábito de compra semanal es el más frecuente entre los clientes de Eroski, seguido por una fracción importante con un comportamiento quincenal. Asimismo, existe una base de consumidores que realiza compras frecuentes de corta duración, probablemente asociadas a necesidades puntuales o productos frescos. Estos hallazgos permiten concluir que la planificación comercial y las campañas promocionales pueden beneficiarse significativamente al considerar estos ciclos de compra.

### Clustering de Clientes

**Informe de Análisis de Clustering para Segmentación de Clientes**

1. **Introducción**

Este informe presenta el proceso y resultados del análisis de clustering realizado sobre datos de clientes. El objetivo del análisis fue identificar segmentos naturales de clientes que permitan entender patrones de comportamiento y características similares para una mejor toma de decisiones estratégicas.

1. **Preparación de Datos**
   1. **Detección y eliminación de outliers**

Para garantizar que los clusters identificados no estén influenciados por observaciones extremas, se aplicó un proceso de detección y eliminación de outliers:

* Se utilizó el método IQR (Rango Intercuartílico) para identificar valores atípicos
* Se aplicó un coeficiente de 1.5 para determinar los límites de detección
* Se eliminaron todas las observaciones identificadas como outliers en cualquiera de las variables numéricas

Resultados de la detección de outliers:

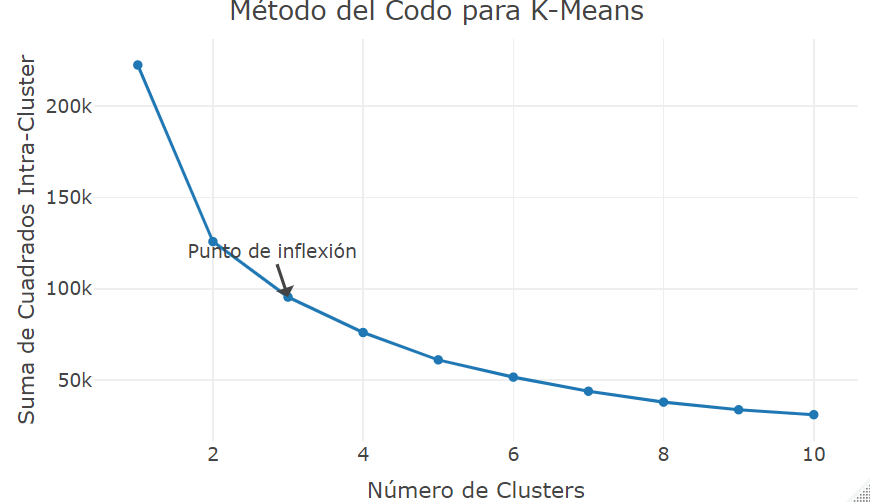
* Se detectó un porcentaje significativo de outliers que fueron eliminados para mejorar la calidad del clustering
* Las dimensiones del conjunto de datos se redujeron, manteniendo observaciones más representativas
  1. **Escalado de variables**

Previo al análisis de clustering, todas las variables numéricas fueron estandarizadas mediante la función scale() de R para evitar que variables con rangos más amplios dominaran el proceso de agrupación.

**2.3 Selección del Número Óptimo de Clusters**

Para determinar el número óptimo de clusters, se aplicó el método del codo (Elbow Method):

* Se calculó la suma de cuadrados intra-cluster (WSS) para valores de k entre 1 y 10
* El análisis del gráfico del método del codo mostró un punto de inflexión en k=3
* Se seleccionó k=3 como el número óptimo de clusters para ambos métodos de clustering



**2.4 Métodos de Clustering Aplicados**

Se implementaron y compararon dos métodos de clustering:

1. **K-means**:
   * Algoritmo particional que minimiza la suma de distancias al cuadrado
   * Se ejecutó con k=3 y múltiples inicializaciones aleatorias (nstart=25)
   * Se calculó el coeficiente de silueta para evaluar la calidad de los clusters
2. **Clustering Jerárquico**:
   * Se utilizó el método de Ward (ward.D2) para minimizar la varianza intra-cluster
   * Se cortó el dendrograma para obtener 3 clusters
   * Se calculó el coeficiente de silueta para comparar con K-means

**3. Resultados**

**3.1 Comparación de Métodos**

Los métodos fueron evaluados utilizando el coeficiente de silueta, que mide la cohesión y separación de los clusters:

|  |  |
| --- | --- |
| **método** | **Coeficiente de Silueta Media** |
| k-means | 0.3570239 |
| Jerarquico | 0.3029755 |

**3.2 Visualización de Clusters**

Se aplicó Análisis de Componentes Principales (PCA) para reducir las dimensiones y visualizar los clusters:

* Los primeros dos componentes principales fueron utilizados para crear gráficos de dispersión
* Los clusters resultantes de ambos métodos mostraron patrones de agrupación claramente diferenciados
* Se observó una distribución espacial similar en ambos métodos, lo que refuerza la consistencia del análisis

**3.3 Caracterización de Clusters**

1. **Perfiles de Clusters K-means**

A partir del gráfico de radar para K-means, se identificaron los siguientes perfiles:

**Cluster 1 (rojo)**:  
Clientes con valores bajos en todas las variables. Son usuarios poco activos, con pocas compras y variedad limitada de productos. Representan el grupo de menor interacción con el sistema.

**Cluster 2 (naranja)**:  
Usuarios con un comportamiento medio. Tienen una actividad y variedad de productos algo superior al Cluster 1, pero sin llegar a los niveles del Cluster 3. Se trata de clientes regulares.

**Cluster 3 (verde)**:  
Este grupo destaca claramente en todas las variables. Son los más activos, con más productos distintos y mayor número de compras entre semana y en total. Representan a los clientes más valiosos por su alta frecuencia de compra y diversidad.

Aquí un anexo, para acceder al grafico de radar de perfiles de clusters para k-means:

PERFIL CLUSTER K-MEANS

**Perfiles de Clusters Jerárquicos**

Según el gráfico de radar del clustering jerárquico, los perfiles son:

**Cluster 1 (rojo)**:  
Clientes de baja actividad, similares al Cluster 1 de K-means. Presentan pocos días activos y baja variedad de productos.

**Cluster 2 (naranja)**:  
Grupo con una frecuencia de compra semanal algo mayor que el Cluster 1, pero con un consumo general bajo. Puede representar clientes más constantes, pero de bajo gasto.

**Cluster 3 (verde)**:  
Este clúster tiene los valores más altos en casi todas las variables, especialmente en total de productos y compras entre semana. Representa a los clientes más intensivos y diversos en sus hábitos de compra.

Aquí un anexo, para acceder al grafico de radar de perfiles de clusters para jerarquico:

PERFIL CLUSTER JERARQUICO

**3.4 Distribución de Variables por Cluster**

Se analizó la distribución de las variables más importantes por cluster mediante gráficos de caja (boxplots)

**4. Conclusiones**

* El análisis de clustering permitió identificar 3 segmentos de clientes bien diferenciados
* Ambos métodos (K-means y Jerárquico) mostraron resultados consistentes, con k-means presentando un mayor coeficiente de silueta
* Los perfiles de clusters obtenidos proporcionan información valiosa para estrategias personalizadas de marketing y servicio al cliente

1. **5. Recomendaciones**

* Implementar estrategias de marketing diferenciadas para cada segmento identificado
* Realizar un seguimiento periódico de estos segmentos para detectar cambios en el comportamiento
* Considerar la incorporación de variables adicionales en futuros análisis para enriquecer la caracterización de los segmentos

1. **Anexos**

* Gráficos de radar de centroides
* Gráficos de dispersión de clusters
* Boxplots de variables importantes por cluster

### Modelización

*Descripción de los métodos y modelos utilizados para la construcción de los diferentes sistemas de recomendación*

*Obtención de resultados de los 4 objetivos-problemáticas propuestos*

*Visualización de resultados*

### Conclusiones

*En un par de párrafos o tres resumir qué se concluye de todo el análisis y modelos de los datos.*

## Transformar los negocios

En este proyecto hemos diseñado un sistema de recomendación para los clientes de EROSKI en su tienda online, basándonos en sus historiales de compra. Nuestra idea era simple pero potente: conseguir que cada cliente reciba sugerencias de productos que realmente le puedan interesar.

Comenzamos analizando los datos de compras para entender mejor cómo se comportan los clientes. Revisamos qué productos suelen comprar juntos, cuáles son sus marcas favoritas y cómo cambian sus hábitos según la temporada. Este análisis nos dio una visión clara de los patrones de consumo que podíamos aprovechar.

Después, probamos diferentes técnicas para hacer recomendaciones. Utilizamos filtrado colaborativo, que básicamente sugiere productos que han comprado personas con gustos parecidos. También implementamos modelos basados en contenido, que recomiendan productos similares a los que el cliente ya ha comprado antes. Experimentamos tanto con métodos tradicionales como con algoritmos más avanzados para ver cuáles funcionaban mejor.

A través de este proyecto, hemos aprendido mucho sobre cómo usar los datos para personalizar la experiencia de compra. Hemos visto que un buen sistema de recomendación puede hacer que los clientes se sientan más conectados con la tienda y descubran productos que realmente les gustan.

También nos ha servido para entender cómo funcionan estos sistemas en un negocio real y no solo en teoría. Hemos comprobado que la personalización puede transformar completamente la relación entre una empresa y sus clientes, haciendo que la experiencia de compra sea más agradable y relevante para cada persona.

## Implicaciones legales y éticas

El uso de datos por parte de Eroski para personalizar la experiencia de compra plantea importantes implicaciones éticas que merecen una reflexión profunda. Aunque la recolección de datos a través de transacciones en línea y tarjetas de fidelización es técnicamente válida, es esencial que los clientes sean informados de manera clara, accesible y comprensible sobre qué datos se están recopilando, cómo se procesan y con qué fines específicos. La transparencia y el consentimiento informado deben ser principios fundamentales, no meras formalidades legales.

Una cuestión clave es si los consumidores realmente tienen control sobre el uso de sus datos. Sería éticamente deseable permitir que los clientes opten por no permitir el uso de sus datos para ciertas aplicaciones, como los sistemas de recomendación, sin que esto afecte negativamente su experiencia de compra. Esta opción no solo respeta la autonomía individual, sino que también se alinea con los principios de protección de datos del Reglamento General de Protección de Datos (RGPD) y el derecho a la autodeterminación informativa.

Desde la perspectiva de la misión de Eroski, que promueve un consumo responsable, es fundamental que las recomendaciones de productos se alinean con estos valores. Las sugerencias podrían incluir etiquetas o indicaciones sobre el impacto ambiental, la estacionalidad o el origen local de los productos, para fomentar un consumo más consciente. De esta manera, además de contribuir al ODS 12 (Producción y Consumo Responsables), el proyecto podría tener efectos positivos en otros ODS como el 3 (Salud y Bienestar) y el 13 (Acción por el Clima), al promover una alimentación saludable y sostenible.

Sin embargo, es crucial reflexionar sobre el verdadero objetivo del proyecto: ¿se busca fomentar el ahorro y una mejor planificación del consumo, o simplemente aumentar las ventas a través de recomendaciones más efectivas? La línea entre ofrecer un servicio realmente útil y fomentar un consumo impulsivo puede ser bastante delgada, especialmente cuando se emplean sesgos cognitivos como la escasez (“solo hoy”), el anclaje (“antes a 5€, ahora a 3€”) o el arrastre social (“otros también han comprado…”). Aunque estas estrategias son bastante comunes en el comercio, su uso excesivo o sin una justificación clara puede chocar con el código ético de Eroski y poner en riesgo su compromiso con un consumo responsable.

Desde un punto de vista técnico, también es importante considerar las limitaciones que vienen con los datos disponibles. Si los datos están sesgados —por ejemplo, si se recopila más información de ciertos perfiles de clientes o de determinadas regiones— esto puede llevar a recomendaciones que no sean representativas o justas. Además, los modelos de segmentación de clientes no deberían usar etiquetas como “cliente responsable” o “irresponsable”, ya que esto puede resultar en juicios de valor injustos o discriminatorios.

En cuanto al impacto social, sería genial comunicar este tipo de iniciativas a través de la revista Consumer, siempre con un enfoque claro y transparente. Explicar cómo se analizan los datos y qué beneficios se buscan puede ayudar a fortalecer la confianza de los clientes y fomentar una mayor participación social en la transformación del consumo. También sería útil establecer canales donde las personas puedan corregir errores en sus datos o expresar su desacuerdo con ciertas segmentaciones o recomendaciones.

Finalmente, desde la perspectiva del Data Ethics Canvas, hay algunos elementos clave que destacar:

* Fuentes de datos: los datos se recopilan a partir de las propias transacciones de los clientes. Al incluir datos potencialmente sensibles, como hábitos de consumo o productos específicos, se necesita un tratamiento ético más riguroso.
* Impacto positivo: el proyecto puede ayudar a las familias a gestionar mejor su gasto y reducir el desperdicio, lo que tendría un efecto favorable en los ODS mencionados.
* Impacto negativo: el riesgo de manipular decisiones de compra mediante técnicas persuasivas puede ir en contra de la voluntad del consumidor y del compromiso ético de Eroski.
* Transparencia y participación: sería deseable implementar mecanismos para que las personas usuarias puedan conocer cómo se usan sus datos, corregir errores y tener voz sobre su uso.
* Plan de acción: antes de avanzar, se debería definir una política clara de consentimiento, revisar los algoritmos en busca de sesgos, y establecer canales de comunicación y educación para la clientela sobre el uso de sus datos.

## Bibliografía

*De artículos científicos y fuentes contrastadas (Wikipedia no lo es)*

## ANEXOS

Gráfico, Gráfico radial

El contenido generado por IA puede ser incorrecto.Gráfico, Gráfico radial

El contenido generado por IA puede ser incorrecto.

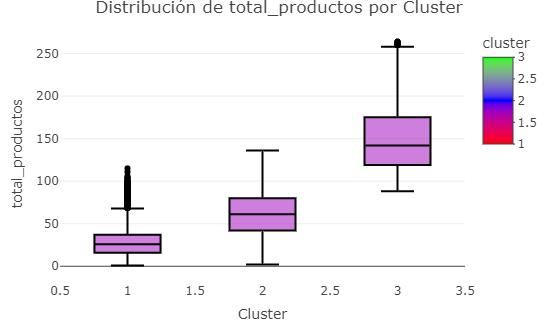
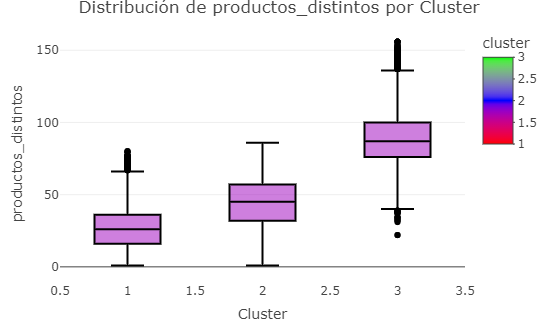
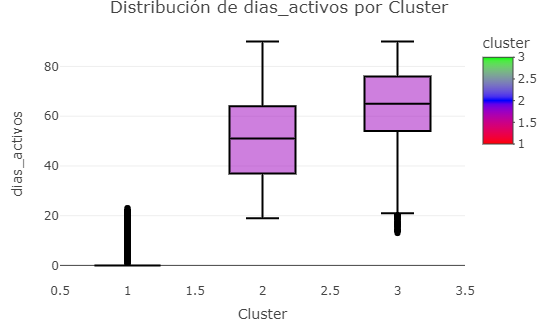
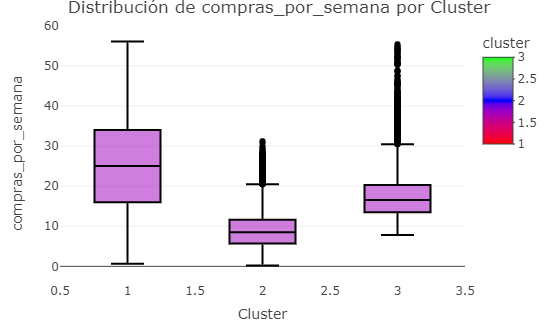
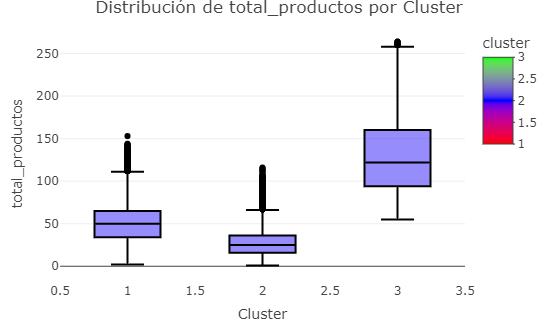
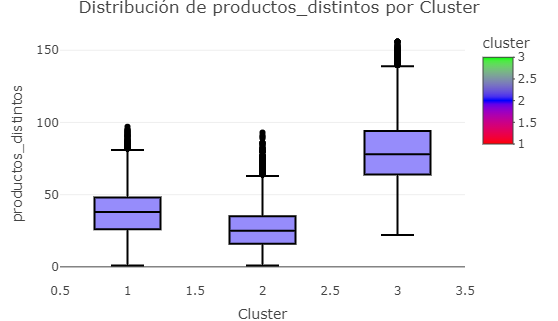
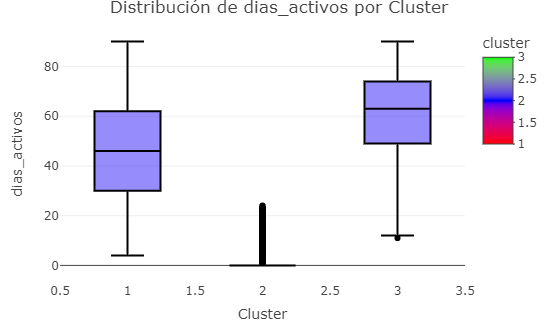
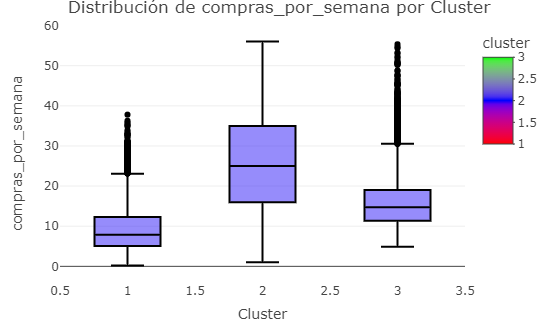
GRAFICOS DE DISPERSION K-MEANS Y JERARQUICO

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.

GRAFICOS BOXPLOT DE K-MEANS Y JERARQUICO



**WRMF Y ALS EXPLICACION**

En este proyecto se implementa un sistema de recomendación basado en **Weighted Regularized Matrix Factorization (WRMF)**, una técnica matemática que permite descomponer la matriz de interacciones entre usuarios y productos en factores latentes. Para resolver esta descomposición se utiliza el algoritmo **Alternating Least Squares (ALS)**, que optimiza iterativamente los vectores de usuarios e ítems mediante mínimos cuadrados alternos.

Esta metodología combina conceptos de álgebra lineal y optimización, permitiendo generar recomendaciones personalizadas a partir de interacciones implícitas. A continuación, se expone la base matemática del modelo y su aplicación al problema propuesto.

**WRMF:**

Weighted Regularized Matrix Factorization, conocido como WRMF, es una técnica de factorización matricial diseñada específicamente para abordar el problema de recomendación, es decir, cuando no se cuenta con calificaciones explícitas de los usuarios, sino con interacciones observadas como clics, compras, visualizaciones o reproducciones. A diferencia de los métodos tradicionales de factorización como SVD que trabajan con calificaciones explícitas, WRMF se adapta mejor a situaciones en las que el sistema solo puede observar si el usuario ha interactuado o no con un ítem, pero no conoce su nivel de satisfacción.

La base de WRMF es la misma que en otros métodos de factorización matricial: descomponer una matriz de interacciones usuario-ítem en dos matrices latentes más pequeñas, una que representa a los usuarios y otra a los ítems, de forma que su producto aproxima la matriz original y permite predecir interacciones faltantes. Sin embargo, en lugar de tratar las interacciones observadas como calificaciones con valores reales, WRMF las trata como preferencias binarias donde un uno indica que hubo interacción y un cero indica ausencia de ella. Este enfoque introduce una diferencia importante en la forma de modelar la confianza en los datos.

Para poder aplicar la factorización correctamente en un contexto tan sesgado, WRMF introduce una matriz de pesos que actúa como un sistema de confianza en cada observación. El algoritmo supone que si un usuario ha interactuado con un ítem, entonces hay una alta probabilidad de que le guste, por lo que esa entrada debe tener más peso en el modelo. Por otro lado, la ausencia de interacción no necesariamente implica que el usuario no esté interesado, por lo que estas entradas también se tienen en cuenta pero con un peso menor. Esta ponderación diferencial permite capturar la incertidumbre propia del feedback implícito.

WRMF es especialmente útil en escenarios reales donde no se cuenta con evaluaciones explícitas sino con registros de comportamiento del usuario. Plataformas como tiendas online, servicios de streaming o motores de búsqueda lo utilizan para generar recomendaciones personalizadas basadas en hábitos de navegación o consumo. Su capacidad para modelar la incertidumbre en la falta de interacción y para asignar diferentes niveles de importancia a los datos observados lo convierte en un método robusto y muy efectivo para construir sistemas de recomendación con feedback implícito.

**ALS**:

**ALS** es un algoritmo utilizado en sistemas de recomendación para la **factorización de matrices**. Su objetivo es **predecir las valoraciones faltantes** en una matriz de usuario-producto (por ejemplo, qué puntuación le daría un usuario a una película que aún no ha visto).

La idea es que una matriz grande de interacciones (como valoraciones de usuarios a productos) puede aproximarse como el producto de **dos matrices de menor rango**.

**Notación**

Supongamos que tenemos:

* Una matriz    
   Valoraciones de m usuarios sobre n productos. La matriz es generalmente dispersa, es decir, la mayoría de las entradas están vacías.
* Buscamos dos matrices:  
    
  +  matriz de características de usuarios
  +  matriz de características de productos
* Tal que:  
   

Donde f es el número de factores latentes, mucho menor que m o n.

**¿Cómo funciona ALS?**

ALS resuelve el problema minimizando la siguiente función de coste:Imagen que contiene objeto, reloj

El contenido generado por IA puede ser incorrecto.

* Ω es el conjunto de pares (i,j) donde hay una valoración.
* λ es el parámetro de regularización, que evita el overfitting.
* Ui es la fila i ésima de la matriz U (usuario i).
* Pj es la fila j ésima de la matriz P (producto j).

**Alternating Least Squares (ALS)**

El nombre "alternating" se debe a que:

1. Se **fija P** y se resuelve para **U** (minimización cuadrática).
2. Luego se **fija U** y se resuelve para **P**.
3. Se alterna entre estos dos pasos hasta la convergencia.

Cada paso involucra resolver una serie de **problemas de mínimos cuadrados**, lo cual es computacionalmente eficiente.

**¿Cómo se usa ALS en el reto?**

Se usa WRMF con ALS para **crear un sistema de recomendación personalizado** basado en interacciones implícitas de los usuarios (clics, visitas, compras). También, **ALS** permite encontrar las matrices latentes X (usuarios) y Y (productos) para luego predecir nuevas recomendaciones.

**Pasos con ALS:**

1. **Preprocesar los datos**:  
     
   * Crear la matriz R con usuarios y productos.
   * Dividir en entrenamiento y validación/test.
2. **Entrenar el modelo ALS**:  
     
   * Elegir hiperparámetros, número de factores f, λ, número de iteraciones.
   * Entrenar alternando entre actualizar U y P.
3. **Predecir las valoraciones faltantes**:  
     
   * Usar  para los pares no observados.
4. **Evaluar el modelo**:  
     
   * Métricas común: RMSE

**Ejemplo visual**

Si un usuario ha puntuado:

|  |  |  |  |
| --- | --- | --- | --- |
| **Usuario** | **Película A** | **Película B** | **Película C** |
| U1 | 5 | ? | 3 |
| U2 | 2 | 4 | ? |
| U3 | ? | 1 | 5 |

Con ALS, aprendes factores latentes de usuarios y películas que permiten predecir los signos de interrogación.