Tecnológico de Monterrey
Escuela de Ingeniería y Ciencias

**Campus Monterrey**

**Desarrollo de aplicaciones avanzadas de ciencias computacionales (Gpo 503)**

**Entregable 0**


Estudiante:

Daniela Ramos García A01174259

Raymundo Guzmán Mata A01284709

Luis Antonio Barajas Ramirez A01235589

Sergio Ortiz Malpica A01284951

Arturo Durán Castillo A00833516


Fecha de entrega:

24 de marzo del 2025

**Objetivo:** Comprender el problema a partir de un análisis preliminar de datos, identificar tendencias y generar ideas para la investigación.

1. Descripción del conjunto de datos o fuentes de información:

- **Origen de los datos.**

El conjunto de datos empleado en este análisis fue obtenido desde la plataforma Kaggle, en el marco de la competencia Instacart Market Basket Analysis. Fue publicado por la empresa Instacart en mayo de 2017 con el objetivo de poner a disposición una muestra real y anonimizada de sus datos para fines de investigación, análisis de patrones de consumo y desarrollo de modelos de aprendizaje automático.

El dataset está compuesto por 6 archivos en formato .csv, estructurados como un conjunto de datos relacional, en el que se describen más de 3.4 millones de órdenes de compra realizadas por aproximadamente 206,000 usuarios. Cada usuario cuenta con un historial de entre 4 y 100 pedidos, lo cual permite analizar secuencias de comportamiento de compra a lo largo del tiempo. A continuación una descripción breve de cada archivo.

1. Orders.cvs: Contiene los pedidos realizados por los usuarios.
2. Products.csv: Lista de productos disponibles.
3. Aisles.csv: Lista de pasillos únicos dentro del supermercado.
4. Departments.csv: Lista de departamentos disponibles.
5. Order_products_prior.csv: Contiene los productos incluidos en los pedidos históricos.
6. Order_products__train.csv: Similar a Order_products_prior.csv, pero contiene los productos incluidos en los pedidos correspondientes al conjunto de entrenamiento.

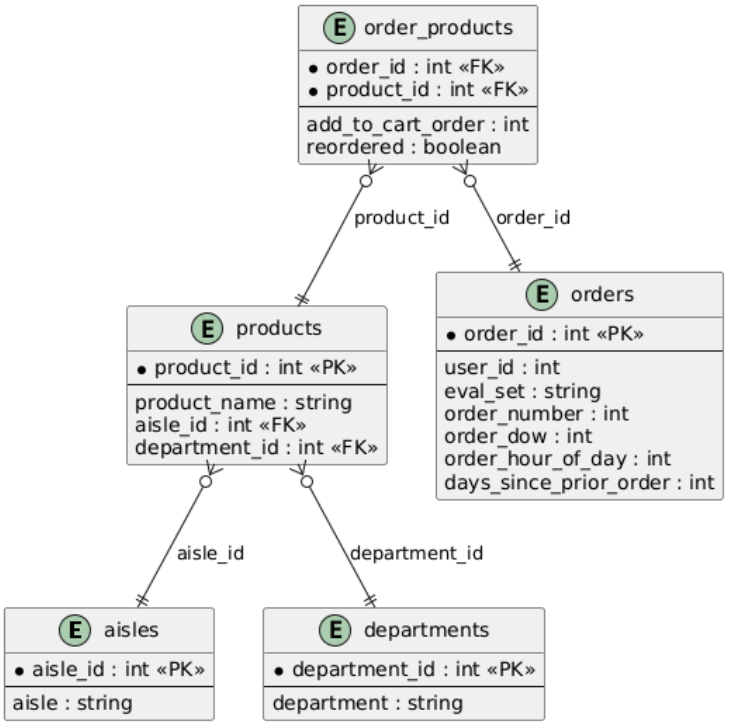
- **Estructura de los datos.**

A continuación, la estructura principal de los datos, incluyendo el número de registros, variables y su tipo general.

1. Orders: 3,421,083 pedidos únicos correspondientes a 206,209 clientes.
2. Products: 49,688 productos únicos.
3. Aisles: 134 pasillos únicos.
4. Departments: 21 departamentos únicos.
5. Order_products: 1,384,617 combinaciones de producto-pedido, 131,209 pedidos únicos y 39,123 productos únicos comprados.

Diagrama relacional ilustrativo del dataset con tipo de datos y variables:

**Figura 1.** Diagrama relacional de los datos




- **Posibles problemas en los datos.**

Se identificaron diversos aspectos que pueden representar problemas en el proceso de limpieza y preparación de los datos para su análisis o modelado. La variable *days_since_prior_order* presenta truncamiento, ya que todos los intervalos mayores o iguales a 30 días se codifican como "30", lo que reduce la granularidad temporal y puede sesgar modelos que dependen de la periodicidad real de compra. Además, esta variable contiene valores nulos para los primeros pedidos de cada usuario, y pueden existir registros incompletos en pedidos cancelados.

Se detectan también inconsistencias en la clasificación de productos por pasillos y departamentos, posiblemente derivadas de errores en la taxonomía o entradas mal etiquetadas. Un hallazgo específico es que 1,258 productos carecen de asignación a departamento, lo cual compromete la integridad relacional de la base de datos. Existe también la posibilidad de registros duplicados, tanto en la tabla de pedidos como en la asociación entre productos y órdenes, lo cual podría impactar el conteo y la evaluación de la recurrencia de compra.

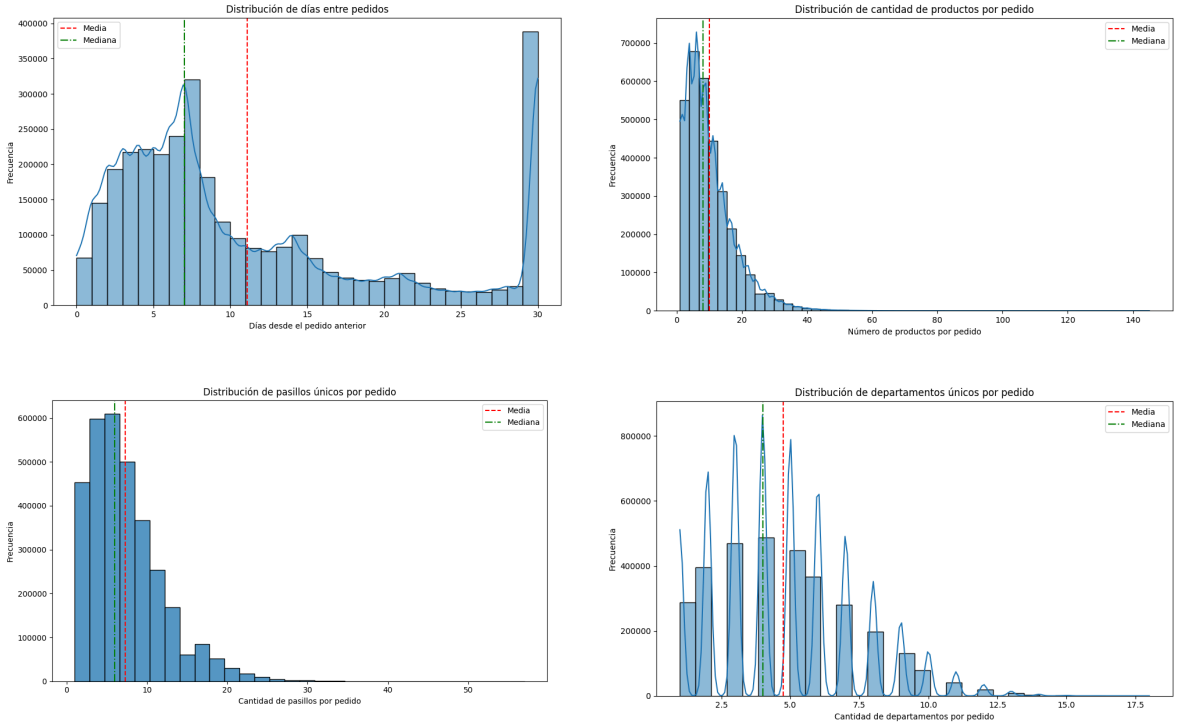
2. Exploración y visualización inicial:

- **Estadísticas descriptivas.**

| **Relaciones** | **Media** | **Mediana** | **Moda** | **Desviación Estándar** |
|---|---|---|---|---|
| Días entre pedidos anteriores. | 11.11 días | 7 días | 30 días | 9.21 |
| Número de productos por orden. | 10.09 productos | 8 productos | 5 productos | 7.5 |
| Número de pasillos incluidos por orden | 7.26 pasillos | 6 pasillos | 5 pasillos | 4.76 |
| Número de departamentos incluidos por orden | 4.73 dpts. | 4.0 dpts. | 4 dpts. | 2.54 |

**Nota:** Los valores mayores a 30 días en la relación de días entre pedidos anteriores han sido truncados, lo que afecta los cálculos de la media, moda y desviación estándar

**Figura 2.** Demostración visual de métricas.




A partir de los datos, podemos observar que la mayoría de personas realiza pedidos aproximadamente cada semana. Sin embargo, existe una alta variabilidad mostrada en la desviación estándar (9.21 días). Si bien, predominan las compras semanales, existe un segmento de clientes que realizan pedidos de forma esporádica.

Las órdenes promedio contienen alrededor de 10 productos distintos, pero la mayoría de pedidos

son más pequeños (5 productos). Esta diferencia entre la media y la moda nos demuestra que si bien predominan los pedidos pequeños, existen órdenes con tamaños muy grandes. Esto se puede observar también a partir de la alta desviación estándar (7.5 productos).

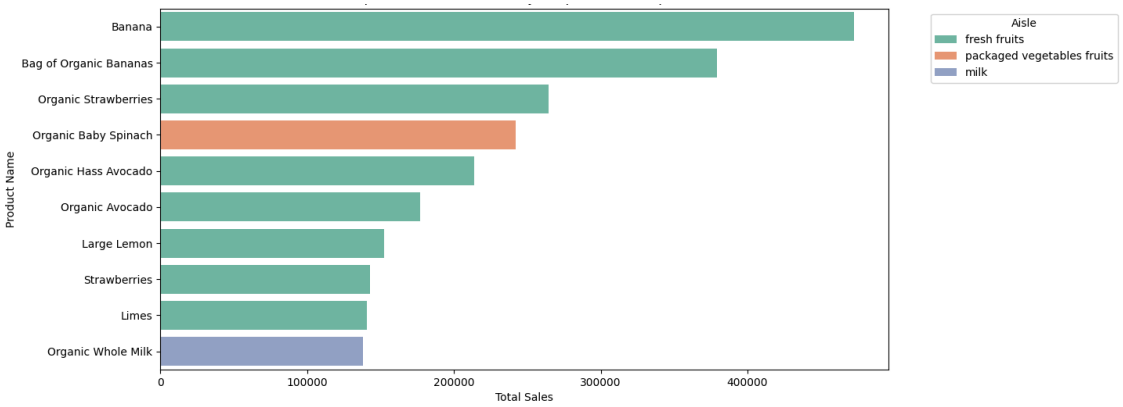
Un pedido típico incluye productos de alrededor de 7 pasillos diferentes y 5 departamentos distintos. Se observa entonces que las compras generalmente abarcan múltiples áreas o categorías, por lo que la compra conjunta de productos pertenecientes a distintos departamentos y pasillos es algo común.

A partir de los datos se puede deducir que actualmente existe una gran oportunidad para optimizar la experiencia del cliente al reorganizar productos considerando la alta variabilidad existente en la cantidad de productos y departamentos por pedido.

- **Visualización de patrones con gráficos adecuados.**

El determinar qué productos específicos generan el mayor volumen de ventas y en qué pasillos se encuentran, nos permite centrarnos en el objetivo principal de optimizar las estrategias comerciales. Al identificar que las frutas frescas, particularmente bananas y fresas orgánicas, dominan las ventas y se concentran principalmente en el pasillo de fresh fruits, podemos tomar decisiones informadas sobre gestión de inventario, distribución de espacios en tienda y estrategias de promoción. Los resultados no solo confirman la fuerte preferencia por productos frescos y orgánicos, sino que también nos permiten focalizar mejor los recursos en las categorías que realmente impulsan las ventas, asegurando que nuestra oferta se alinee con las demandas clave del consumidor.

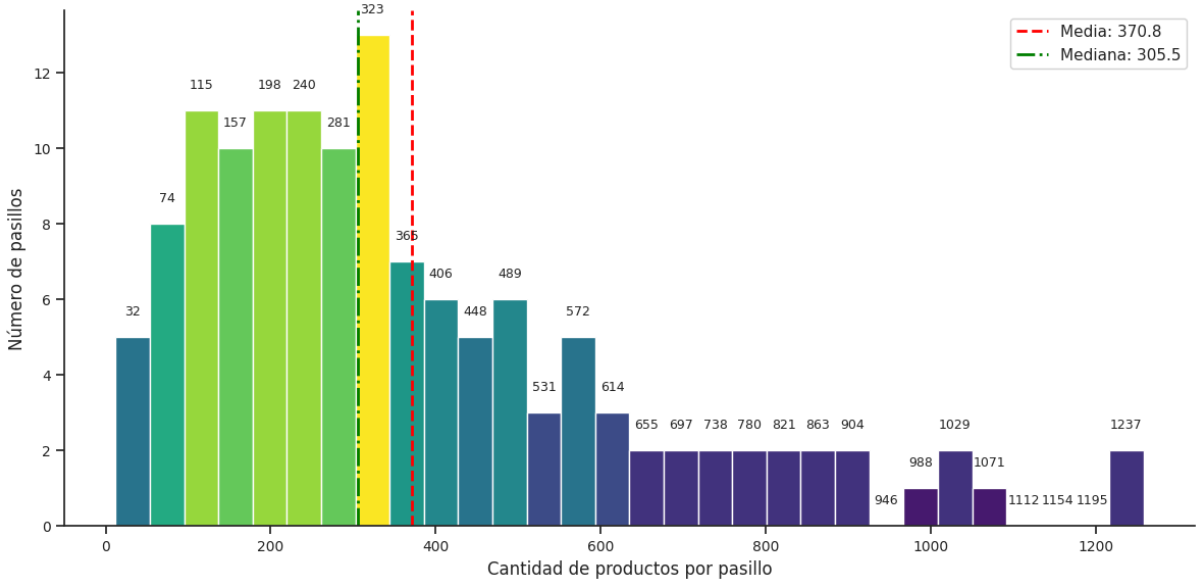
**Figura 3.** Los 10 productos más comprados y con qué pasillo se relacionan.




La distribución actual de productos por pasillo muestra una fuerte desigualdad, con algunos pasillos conteniendo solo unos pocos productos y otros acumulando más de 1,200 (Figura 2).

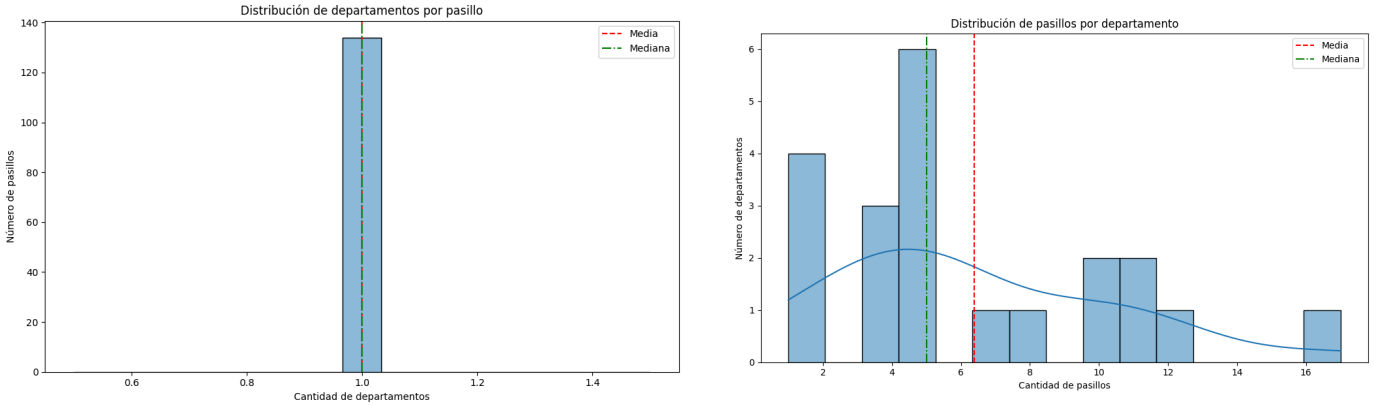
Esta variabilidad indica una oportunidad clara de mejora: redistribuir los productos de manera más balanceada entre pasillos podría facilitar la navegación, reducir la saturación en zonas específicas y mejorar la eficiencia tanto para el cliente como para la operación. Un acomodo basado en afinidad de productos y volumen podría ofrecer una estructura más práctica y funcional.

**Figura 4.** Distribución de productos por pasillo




Analizando la Figura 3 observamos que existe únicamente un departamento en cada pasillo. Partiendo que existen 134 pasillos y 21 departamentos, nos damos cuenta que existen muchos pasillos por cada departamento.

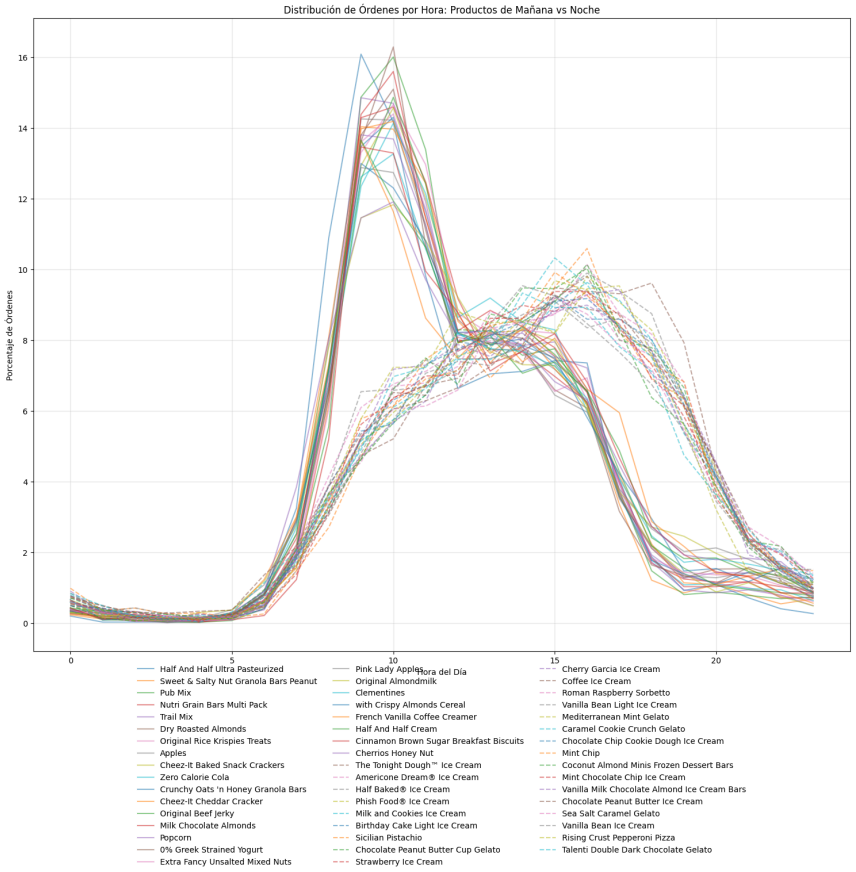
**Figura 5.** Departamentos por pasillo y pasillos por departamento




Los productos más ordenados cambian dependiendo del tiempo en el día. Esto nos ayuda a poder entender más al usuario y su ánimo dependiendo del momento del día. En la Figura 4 podemos

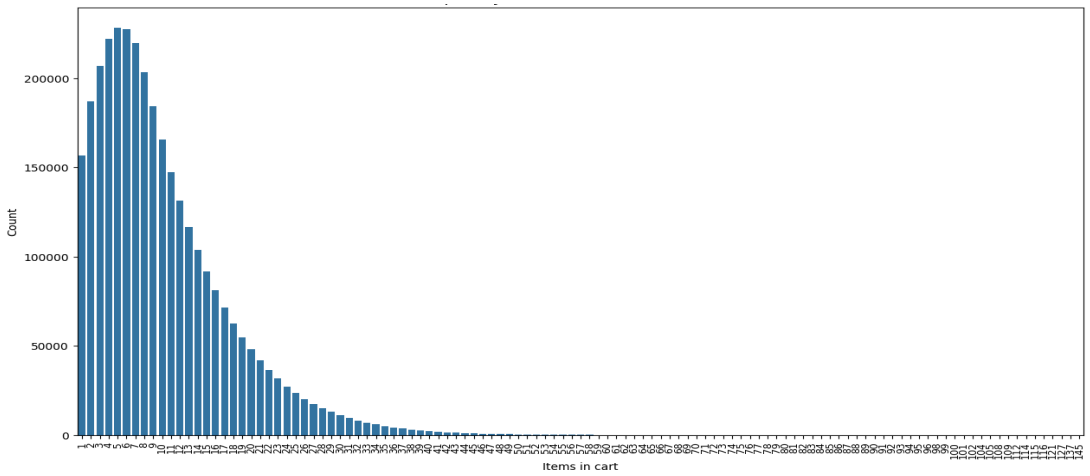
observar que el usuario preferiría productos más acorde a la mañana, tarde o noche y esto representa una tendencia en todos los usuarios. Nos permite mejorar la organización de productos para satisfacer las necesidades del cliente dependiendo del momento del día: si es de mañana es más probable que necesite cosas más 'saludables' como leche 'half and half ultra pasteurized', durante la tarde se presenta una conjetura entre todos los productos y en la noche se piden más 'postres' o alimentos 'express'. No tendría sentido ofrecer productos que no se necesitan durante un tiempo dado en el día si bien sabemos que no se ordenan en ese momento.

**Figura 6.** Distribución de Órdenes por hora



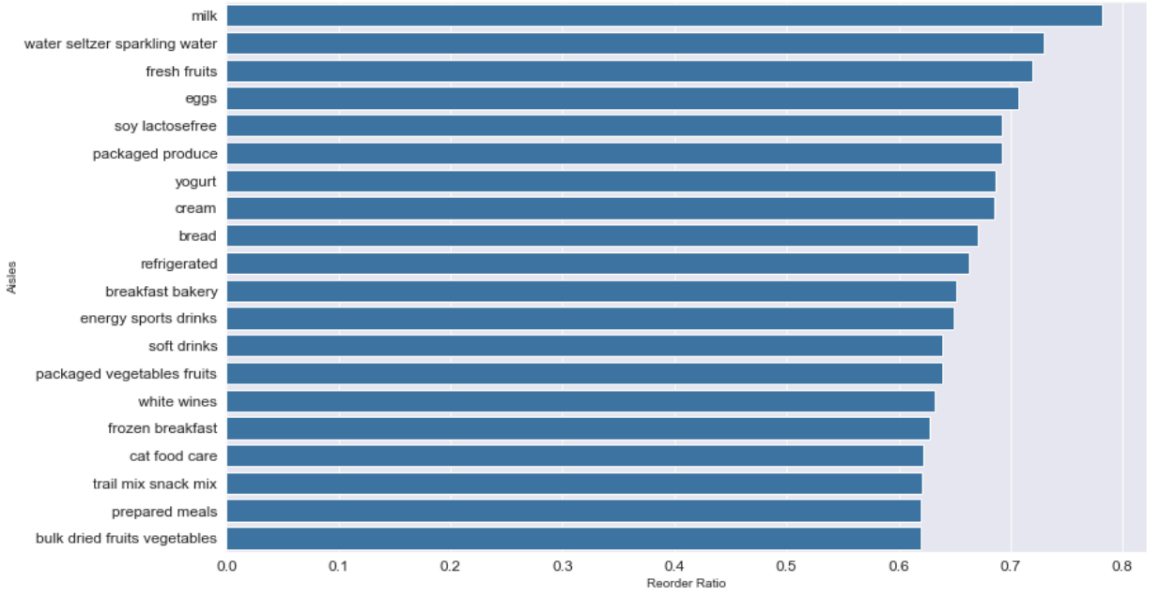

Los usuarios normalmente realizan pedidos 'pequeños', es decir, que sus órdenes están hechas de 1 a 15 productos por orden. En la Figura 5 podemos observar que los usuarios no suelen hacer todo el supermercado si no pedidos por partes 'express' y no definitivas. Podríamos usar este conocimiento para ajustar el orden de los productos para poner productos que 1) son los que más reordena el cliente, 2) son los más comúnmente ordenados en conjuntos (si se ordena leche, probablemente ordenará también huevos, mantequilla y jamón) y 3) más se ordenan en el tiempo del dia.

**Figura 7.** Frecuencia de Productos en Carro




La figura 6 nos permite identificar los pasillos con un mayor ratio de orden. En la misma figura se puede observar que los usuarios suelen frecuentar altamente los pasillos de leche, agua con gas y frutas frescas. Dicha información es importante para poder acomodar los pasillos más frecuentados de forma próxima para mayor comodidad del usuario.

**Figuras 8.** Pasillos con mayor ratio de reorden

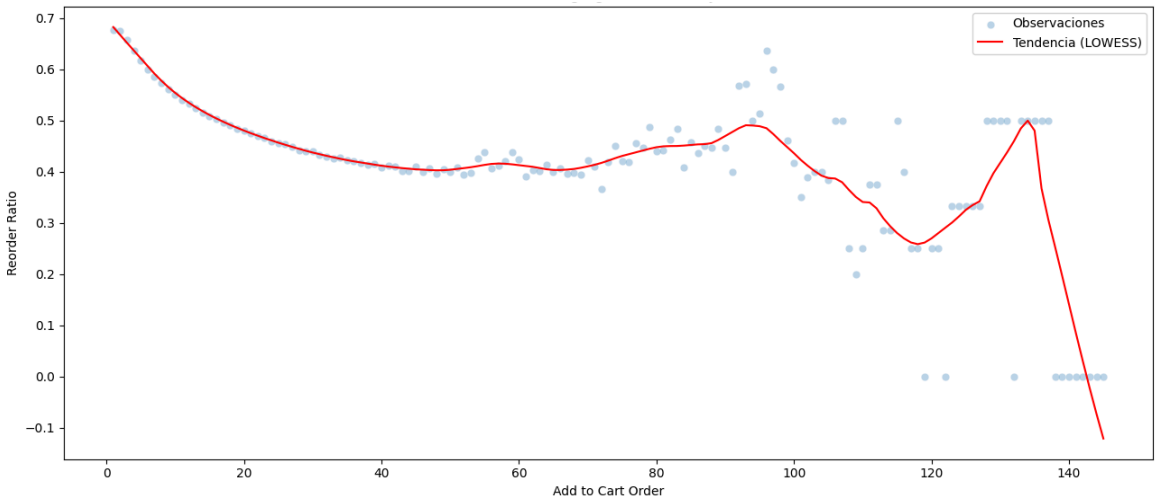



Se tiene una relación inversa entre el orden de agregado al carrito y la tasa de reorden, es decir, los productos seleccionados al inicio tienen una mayor probabilidad de ser reordenados (70%), lo que sugiere que son artículos recurrentes o de alta prioridad para el usuario. A partir de la posición 20 en adelante, la tasa desciende gradualmente y se estabiliza alrededor del 40%,

mientras que en órdenes superiores a la posición 90 se observa mayor variabilidad y ruido, probablemente debido a menor volumen de datos.

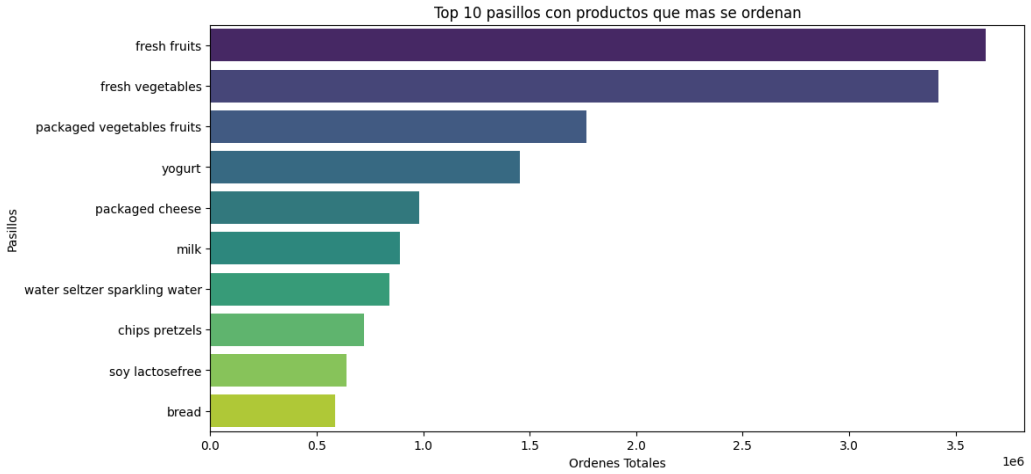
Esto sugiere que los productos frecuentes tienden a ser agregados primero, lo cual puede aprovecharse para reordenar productos en la plataforma priorizando los de mayor recurrencia, optimizando así el flujo de compra.

**Figura 9**. Promedio de reorden de producto dependiendo de la secuencia de selección.




Quisimos analizar los pasillos con mayor demanda de productos para identificar patrones de compra y preferencias clave de los consumidores. Esta gráfica nos permite visualizar de manera clara y rápida los 10 pasillos más populares, destacando que los alimentos frescos como frutas y verduras lideran las ventas, seguidos por productos básicos como lácteos y pan. Este análisis es fundamental para optimizar el manejo de inventario, diseñar estrategias de marketing más efectivas y mejorar la experiencia del cliente al enfocarnos en los productos que generan mayor interés.

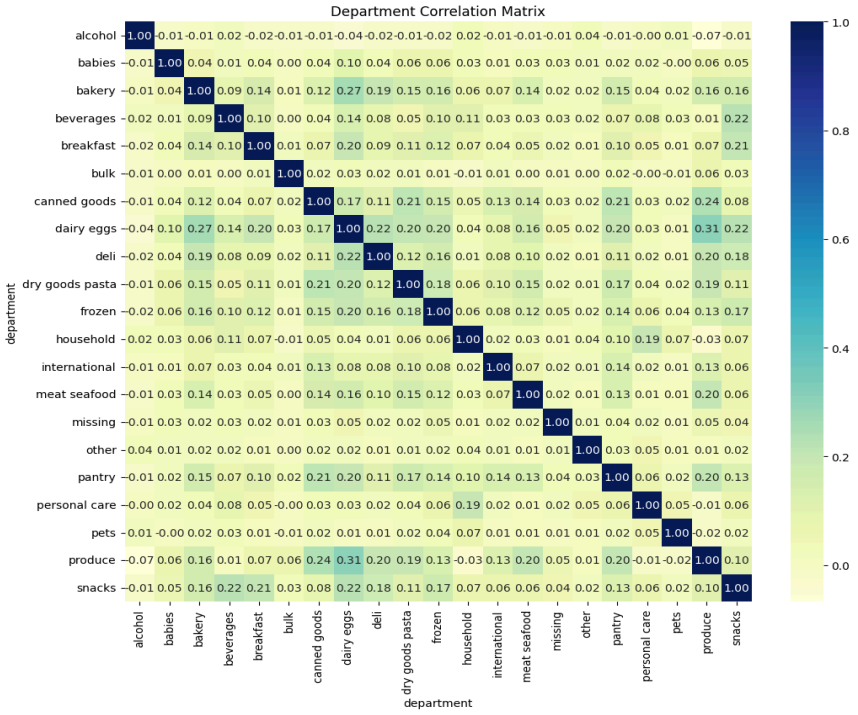
**Figura 10**. Los pasillos con más órdenes de productos



- **Análisis de correlaciones entre variables.**

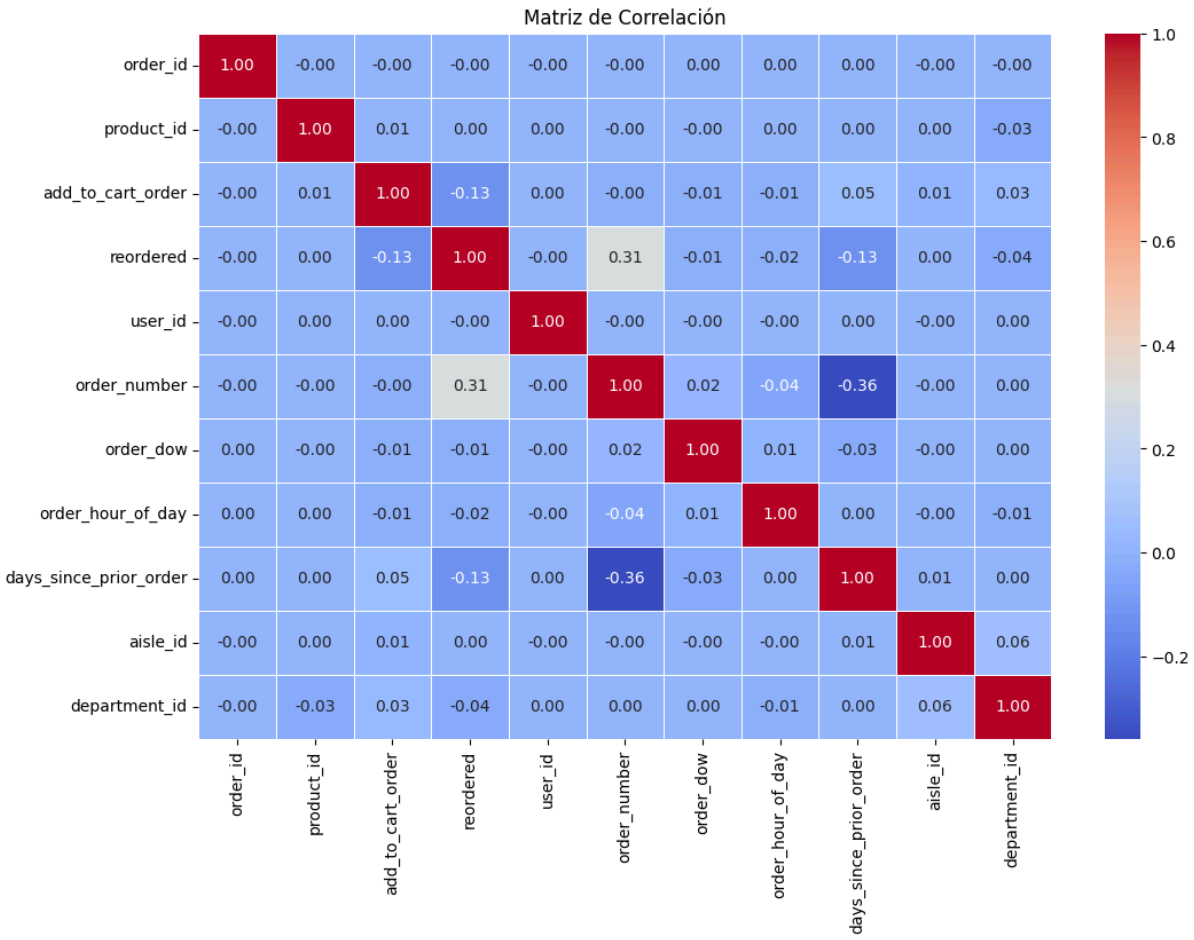
La tabla de correlación entre departamentos nos permite identificar qué departamentos tienen relación en compras. Usemos por ejemplo la relación entre el departamento 'dairy eggs' con 'produce' donde existe una correlación positiva de 0.31, es decir que, si un cliente ordena un producto de 'dairy eggs' muy probablemente también ordene un producto de 'produce'. Esto nos podría permitir acomodar los departamentos que mejor relación tienen entre sí para aumentar las ventas y selección de productos.

**Figura 11.** Matriz de correlación entre departamentos




La matriz de correlación entre variables nos permite observar relaciones clave que revelan patrones de comportamiento en los compradores, por ejemplo el orden de los productos conforme se agregan al carrito (order_number) tienden a reordenar productos con mayor frecuencia (correlación 0.31 con reordered). Estos hallazgos nos permiten entender mejor los hábitos de compra recurrentes y diseñar estrategias más efectivas para fomentar la repetición de compra, optimizando tanto la experiencia del usuario como el potencial de ventas recurrentes.

**Figura 12.** Matriz de correlación entre variables




3. Preguntas preliminares de investigación:

- **Identificación de posibles hipótesis basadas en los datos observados.**

$H_1$: Modificar dinámicamente el orden de los pasillos según los productos más vendidos en diferentes momentos del día disminuirá el tiempo que necesita un usuario para llenar su carrito.

$H_2$: Combinar pasillos que contengan menos de 100 productos con otros que incluyan productos correlacionados disminuirá el número total de pasillos visitados durante una orden de compra.

$H_3$: Agrupar pasillos de diferentes departamentos en una categoría única reducirá el tiempo que necesita el usuario para realizar la selección de sus productos.

$H_4$: Agrupar departamentos con correlación mas fuerte entre ellos reducirá la necesidad del usuario de cambiar de departamento para realizar sus compras.

- **Discusión de enfoques para abordar el problema.**

Creemos que existen 2 posibles enfoques para abordar el problema: enfocado hacia los productos o hacia los pasillos. El primer enfoque como bien lo dice, busca la manera de optimizar el acomodo de los productos por los pasillos así eficientizando las recomendaciones de compras. Ordenar los productos en diferentes pasillos, en donde se acomodarán los productos con mayor relación entre si más cerca uno de otro. Este enfoque se dirige hacia cómo podemos maximizar las relaciones individuales de cada uno de los productos con los demás y viceversa. Es un enfoque más individual, dejando aun lado sus propiedades de departamentos y pasillos a los que pertenecen.

Ahora bien, el segundo enfoque se dirige más hacia los pasillos y departamentos. Se enfoca en un acomodo general de los pasillos en general sin modificar la estructura de jerarquías: un departamento por pasillo. Este busca aprovechar el flujo del cliente durante la compra en línea, es decir, reducir el tiempo de flujo y aumentar ventas. Existen muchas posibles configuraciones de pasillos: juntar pasillos con más correlación de un lado al otro, priorizar los pasillos más importantes por departamentos, reducir cantidad de productos por pasillos, etc. Dado que la solución necesaria es dirigida hacia la reestructuración de los pasillos y productos, creemos firmemente que el segundo enfoque es el más adecuado para investigar.

- **Métricas**

Para evaluar el desempeño del modelo de machine learning en la asignación optimizada de productos a pasillos, se propone utilizar la metrica de *Weighted Reorder Density (WRD)* que evalua la densidad de reorden ponderada por posición, es decir, mide cuántos productos altamente recurrentes están ubicados en posiciones tempranas del carrito.

$$\mathrm{WRD} = \sum_{i=1}^{n} \left( \frac{r_i}{p_i} \right)$$

En un entorno real, la validación podría complementarse con pruebas A/B, comparando la disposición actual con la propuesta, evaluando indicadores como tiempo de recorrido, satisfacción del cliente o eficiencia logística.

<u>• Posibles herramientas a utilizar.</u>

Se realizará el análisis y exploración de datos usando Python como lenguaje principal, trabajando directamente sobre los archivos .csv dentro de un entorno de Jupyter Notebook en Google Colab. Esta elección se basa en la flexibilidad que ofrece para el análisis exploratorio, la facilidad de colaboración en la herramienta, la manipulación de estructuras de datos complejas y la generación de visualizaciones informativas, todo desde una interfaz accesible vía web.

Se hará uso de bibliotecas frecuentemente utilizadas en ciencia de datos, como *pandas* para la gestión eficiente de los datos tabulares, *NumPy* para operaciones numéricas de bajo nivel, herramientas como *matplotlib*, *seaborn* para representar gráficamente los patrones encontrados en los datos y *SciPy* cuando se requieran análisis más detallados o pruebas estadísticas específicas.

Enlace al Jupyter Notebook: organizacion_productos.ipynb

Referencias.


1. Kaggle. (2017). Instacart Market Basket Analysis. Recuperado de https://www.kaggle.com/competitions/instacart-market-basket-analysis
2. Patel, H. (2024). *Cómo realizar pruebas A/B en aprendizaje automático*. Recuperado de https://censius.ai/blogs/how-to-conduct-a-b-testing-in-machine-learning#blogpost-toc-3