

# **Instituto Tecnológico y de Estudios Superiores de Monterrey**



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

Reto Entrega 0

### **Equipo 3 Integrantes:**

Nallely Lizbeth Serna Rivera	A00833111
José Elías Plascencia Cruz	A00832687
Valeria Enríquez Limón	A00832782
Fernando Burgos	A01236284
Antonio A. Chávez Ramos	A00833893

1 de Abril del 2025

## Exploración inicial de datos

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

**Contenido:**

### • Descripción del conjunto de datos o fuentes de información:

**Origen de los datos (dataset disponible, datos generados, recopilación propia, etc.):**

El conjunto de datos proviene de Instacart, un servicio en línea de entrega de comestibles. Instacart ha puesto a disposición una muestra anonimizada de más de 3 millones de órdenes de compra de más de 200,000 usuarios.

**Estructura de los datos: variables, tipos de datos, cantidad de registros.**

Archivos principales:

- orders.csv: Contiene información sobre los pedidos realizados por los usuarios.
- order\_products\_\_prior.csv y order\_products\_\_train.csv: Detalles de los productos en cada pedido, diferenciando entre pedidos anteriores y el conjunto de entrenamiento.
- products.csv: Lista de productos disponibles.
- aisles.csv: Información sobre los pasillos donde se encuentran los productos.
- departments.csv: Datos sobre los departamentos a los que pertenecen los productos.

**Variables y tipos de datos:**

- orders.csv:
  - order\_id (entero): Identificador único del pedido.
  - user\_id (entero): Identificador del usuario que realizó el pedido.
  - eval\_set (cadena): Conjunto al que pertenece el pedido (prior, train, test).
  - order\_number (entero): Número de orden del pedido para el usuario.
  - order\_dow (entero): Día de la semana en que se realizó el pedido.
  - order\_hour\_of\_day (entero): Hora del día en que se realizó el pedido.
  - days\_since\_prior\_order (flotante): Días desde el último pedido.
- order\_products\_\_prior.csv y order\_products\_\_train.csv:
  - order\_id (entero): Identificador del pedido.
  - product\_id (entero): Identificador del producto.
  - add\_to\_cart\_order (entero): Orden en que el producto fue añadido al carrito.

- reordered (entero): Indica si el producto fue reordenado (1) o no (0).
- products.csv:
  - product\_id (entero): Identificador único del producto.
  - product\_name (cadena): Nombre del producto.
  - aisle\_id (entero): Identificador del pasillo.
  - department\_id (entero): Identificador del departamento.
- aisles.csv:
  - aisle\_id (entero): Identificador único del pasillo.
  - aisle (cadena): Nombre del pasillo.
- departments.csv:
  - department\_id (entero): Identificador único del departamento.
  - department (cadena): Nombre del departamento.

#### Cantidad de registros:

- orders.csv: 3,421,083 registros.
- order\_products\_\_prior.csv: 32,434,489 registros.
- order\_products\_\_train.csv: 1,384,617 registros.
- products.csv: 49,688 registros.
- aisles.csv: 134 registros.
- departments.csv: 21 registros.
- Posibles problemas en los datos (valores faltantes, ruido, inconsistencias).
  - En orders la primera orden de cada cliente es un NaN

- Exploración y visualización inicial:

- Estadísticas descriptivas (media, mediana, moda, desviación estándar).

Estadísticas descriptivas:						
	order_id	user_id	order_number	order_dow	\	Moda:
count	3.421083e+06	3.421083e+06	3.421083e+06	3.421083e+06		order_id 1.0
mean	1.710542e+06	1.029782e+05	1.715486e+01	2.776219e+00		user_id 210.0
std	9.875817e+05	5.953372e+04	1.773316e+01	2.046829e+00		order_number 1.0
min	1.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00		order_dow 0.0
25%	8.552715e+05	5.139400e+04	5.000000e+00	1.000000e+00		order_hour_of_day 10.0
50%	1.710542e+06	1.026890e+05	1.100000e+01	3.000000e+00		days_since_prior_order 30.0
75%	2.565812e+06	1.543850e+05	2.300000e+01	5.000000e+00		Name: 0, dtype: float64
max	3.421083e+06	2.062090e+05	1.000000e+02	6.000000e+00		

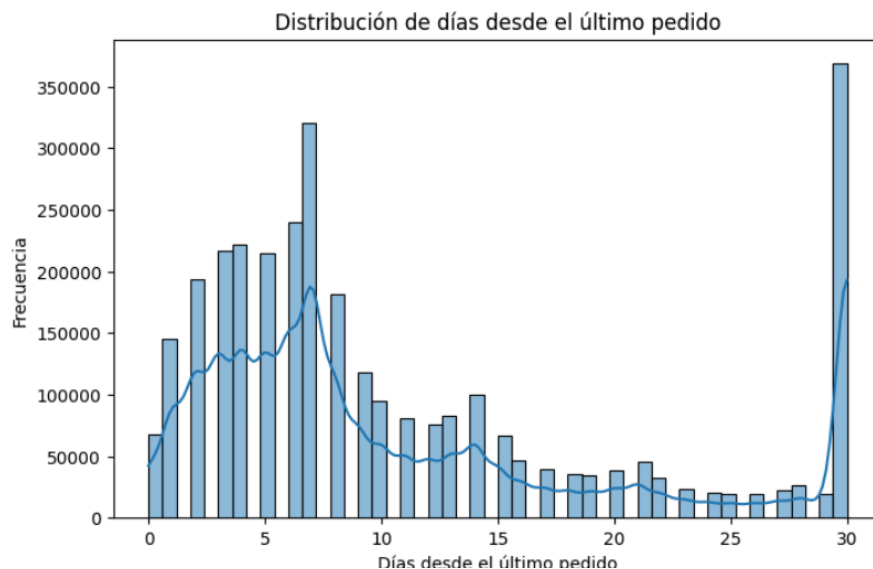
  

order_hour_of_day			days_since_prior_order	
count	3.421083e+06	3.214874e+06		
mean	1.345202e+01	1.111484e+01		
std	4.226088e+00	9.206737e+00		
min	0.000000e+00	0.000000e+00		
25%	1.000000e+01	4.000000e+00		
50%	1.300000e+01	7.000000e+00		
75%	1.600000e+01	1.500000e+01		
max	2.300000e+01	3.000000e+01		

Desviación estándar:	
order_id	987581.739825
user_id	59533.717793
order_number	17.733164
order_dow	2.046829
order_hour_of_day	4.226088
days_since_prior_order	9.206737
dtype:	float64

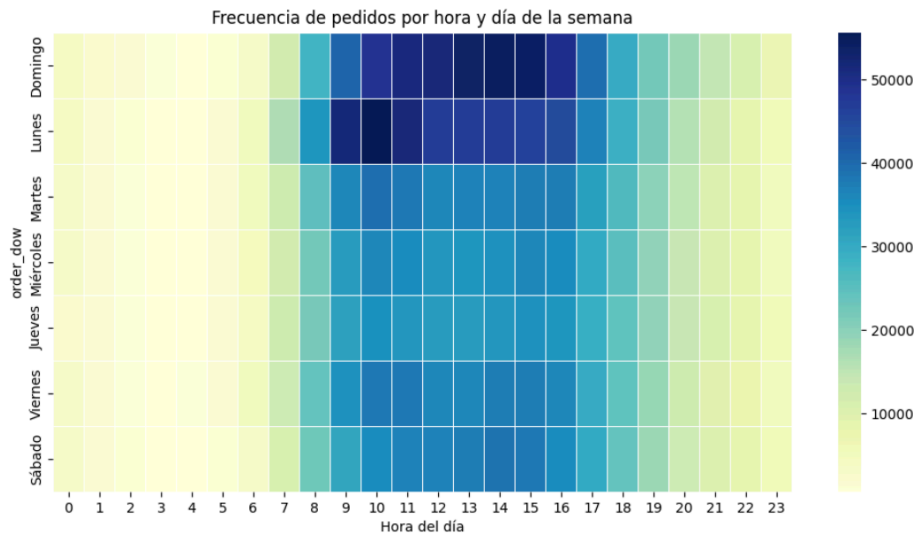
- Visualización de patrones con gráficos adecuados.



La mayoría de los usuarios realizan compras en un rango de 0 a 10 días después de su último pedido, con un pico alrededor del día 7. Esto sugiere que muchas personas tienen un patrón de compra semanal.

Hay otro pico importante en el día 30, lo que indica que algunos clientes hacen pedidos mensuales.

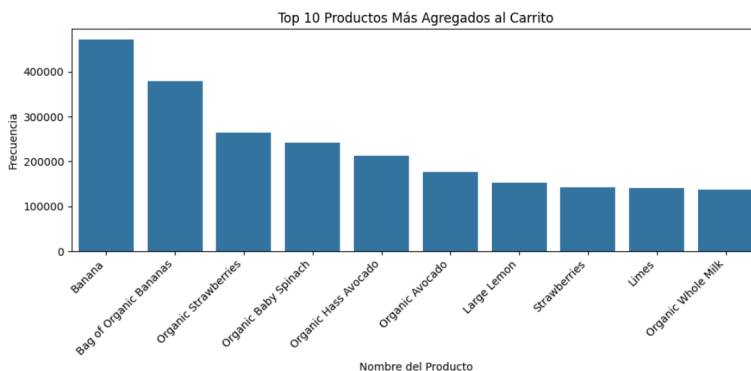
La distribución muestra una caída progresiva después del día 10, lo que sugiere que menos usuarios dejan pasar tanto tiempo entre compras.



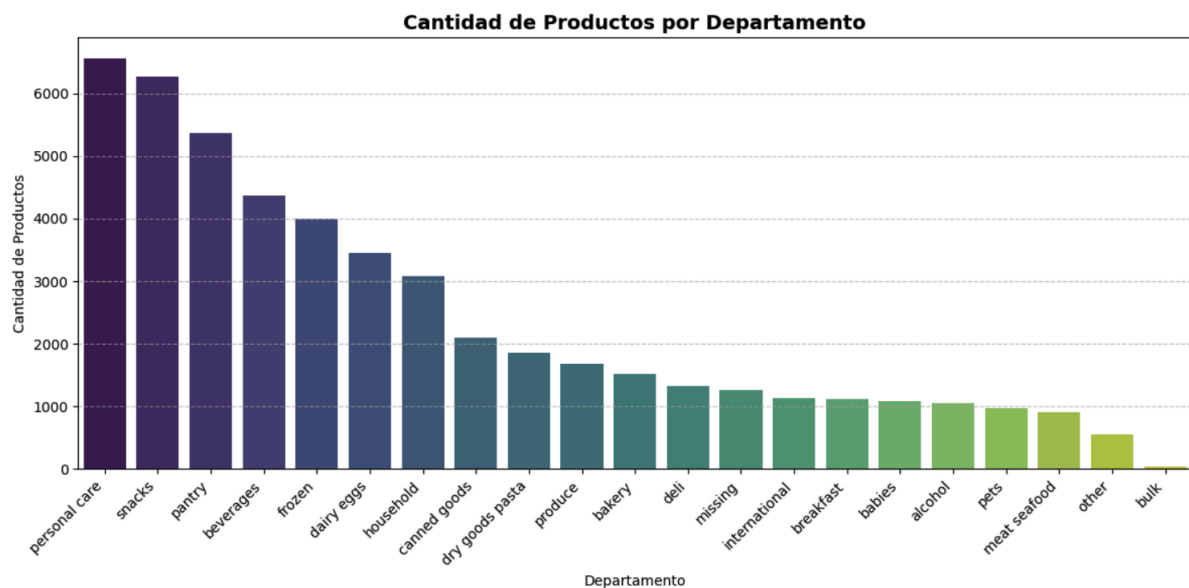
La mayor cantidad de pedidos ocurre entre las 9:00 y las 15:00 horas.

Los días con mayor volumen de pedidos son los fines de semana y los lunes.

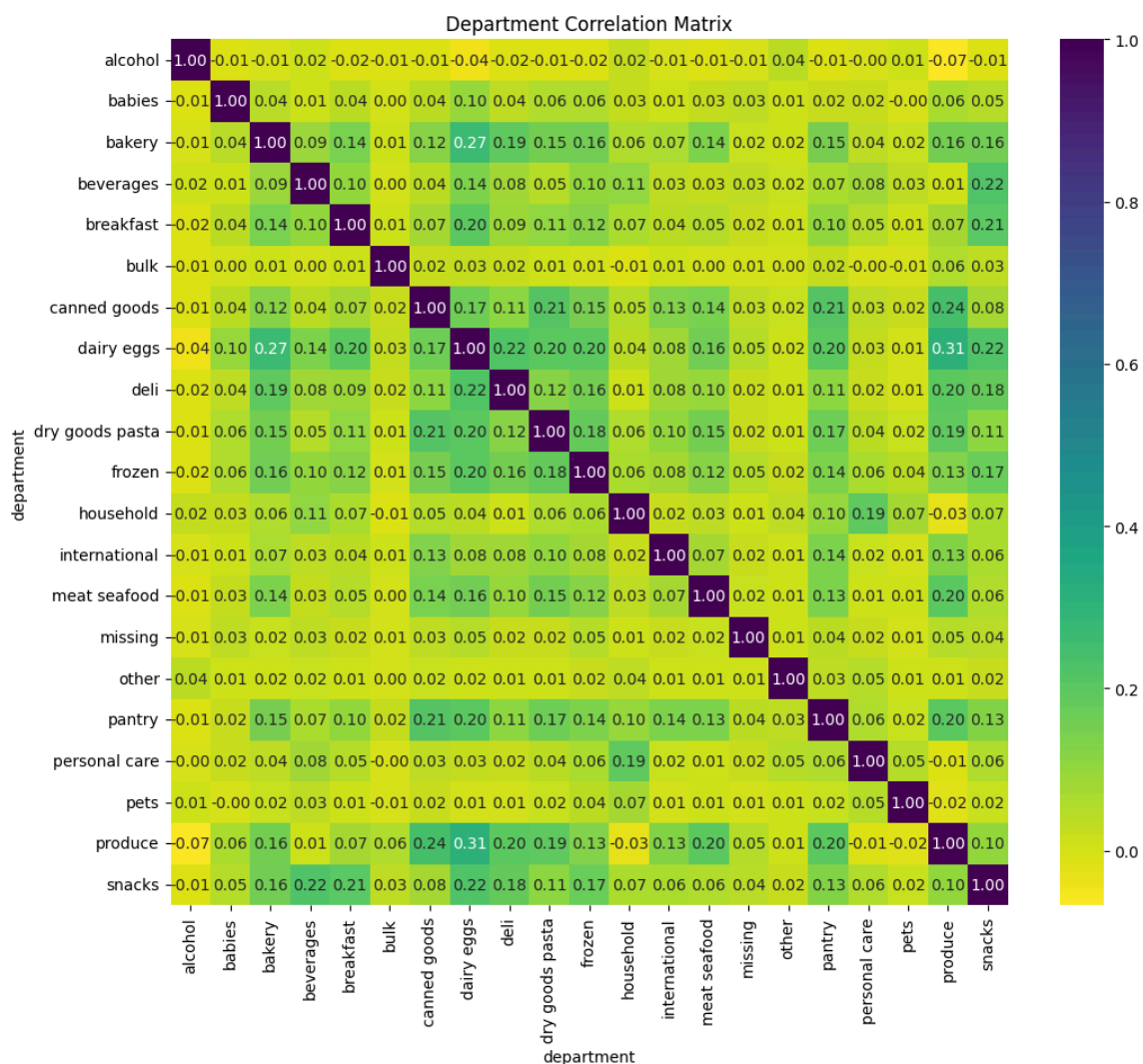
Durante la madrugada (0:00 - 6:00), hay muy poca actividad.



Los productos más comprados son los del departamento de frutas y verduras, abarcando los primeros 9 lugares, y al final se alcanza a observar un producto lácteo. Esto puede ser un indicio de que estos 2 departamentos podrían estar muy relacionados en las compras de los consumidores.



Esta gráfica nos muestra que el departamento de personal care, a pesar de no ser el que más vende, si es el que tiene más productos, por lo que pueden estar quedando productos rezagados. Esto nos dice que no se debe seguir suministrando esos productos ya que puede provocar una pérdida de dinero al invertir y vender dichos productos.



Se observa baja correlación en general, lo que indica que los clientes compran productos de diferentes categorías sin una relación fuerte entre ellos.

Sin embargo, algunas categorías pueden mostrar asociaciones leves, como las frutas y verduras junto con los huevos y lácteos, snacks con bebidas o productos de panadería con desayuno.

- Análisis de correlaciones entre variables.
  - No se encontraron correlaciones significativas en los demás datos.
  - Observamos que el día y hora a la que más se hacen compras es el domingo entre las 8 am y 5 pm
  - La cantidad de productos promedio por carrito es de entre 10 y 12
  - El departamento de personal care es el que más productos tiene, y sin embargo el departamento que más vende es goods
  - Los departamentos que más venden productos en conjunto son “produce” y “dairy eggs”.

Preguntas preliminares de investigación:

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

- Los domingos es el día de la semana cuando más se realizan las compras, a comparación con el resto de la semana.
- Conforme va pasando el mes, se gasta menos, exceptuando el último día del mes, debido al ingreso mensual/quincenal.
- Los productos que más se compran son frutas.
- Las palabras más comunes en los productos son “original”, “gluten free”, “natural” y “organic”.
- El pasillo con más productos es el de “Candy Chocolate”.
- Más de la mitad de la proporción de los productos están “no reordenados”.

Enfoques para abordar el problema.

- **Posibles herramientas a utilizar:** Python (pandas, matplotlib, seaborn), R, SQL, etc.

**Python:**

- **pandas** para manipulación y análisis de datos.
- **matplotlib** y **seaborn** para visualización de datos.
- **Jupyter Notebooks** para documentar y ejecutar el análisis de manera interactiva.
- **NumPy** para cálculos numéricos y manejo de arrays.
- **gdown** para descargar datasets desde Google Drive.
- **pygal** para mejores visualizaciones de gráficos para nuestros resultados.