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 Elias Plascencia Cruz	A00832687
Valeria Limón	A00832782
Fernando Burgos	A01236284

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).

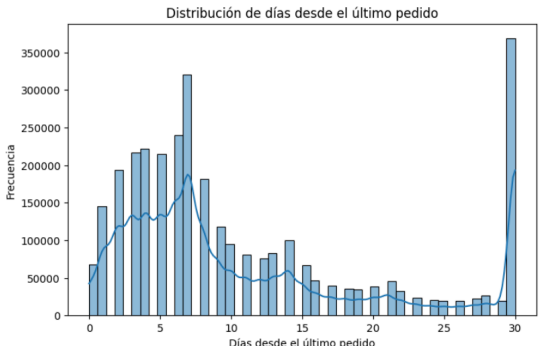
- En orders la primera orden de cada cliente es un NaN

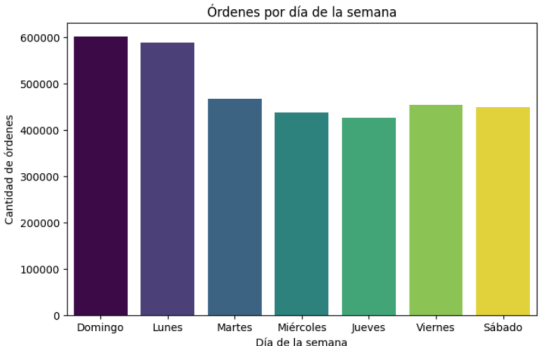
- **Exploración y visualización inicial**:
  - Estadísticas descriptivas (media, mediana, moda, desviación estándar).

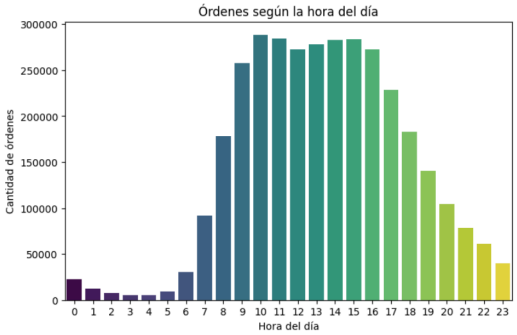
```
Estadísticas descriptivas:
          order_id       user_id  order_number     order_dow  \
count  3.421083e+06  3.421083e+06  3.421083e+06  3.421083e+06
mean   1.710542e+06  1.029782e+05  1.715486e+01  2.776219e+00
std    9.875817e+05  5.953372e+04  1.773316e+01  2.046829e+00
min    1.000000e+00  1.000000e+00  1.000000e+00  0.000000e+00
25%    8.552715e+05  5.139400e+04  5.000000e+00  1.000000e+00
50%    1.710542e+06  1.026890e+05  1.100000e+01  3.000000e+00
75%    2.565812e+06  1.543850e+05  2.300000e+01  5.000000e+00
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
```

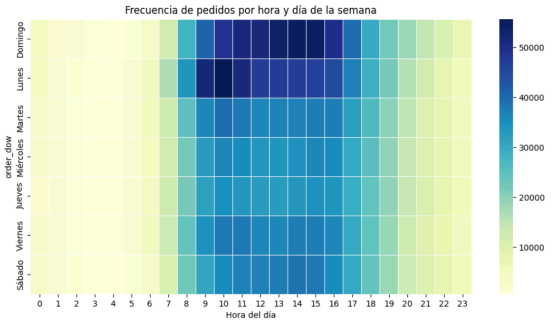
```
Moda:
 order_id                  1.0
user_id                  210.0
order_number               1.0
order_dow                  0.0
order_hour_of_day         10.0
days_since_prior_order    30.0
Name: 0, dtype: float64

Mediana:
 order_id                1710542.0
user_id                  102689.0
order_number                11.0
order_dow                    3.0
order_hour_of_day           13.0
days_since_prior_order       7.0
dtype: float64

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
```

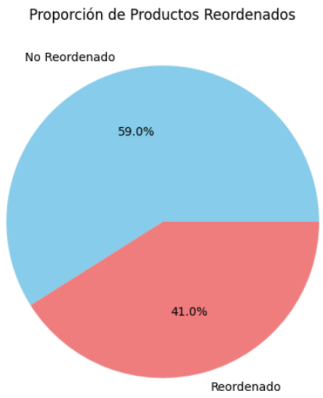
  - Visualización de patrones con gráficos adecuados.

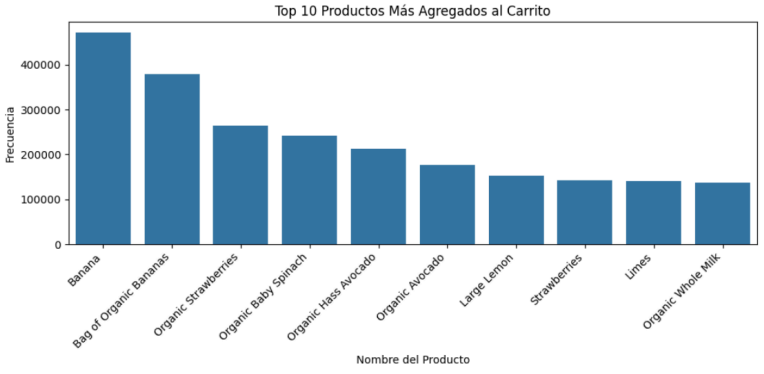


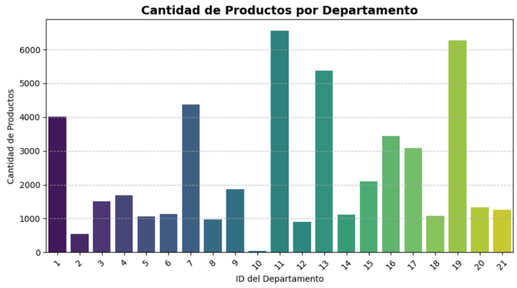


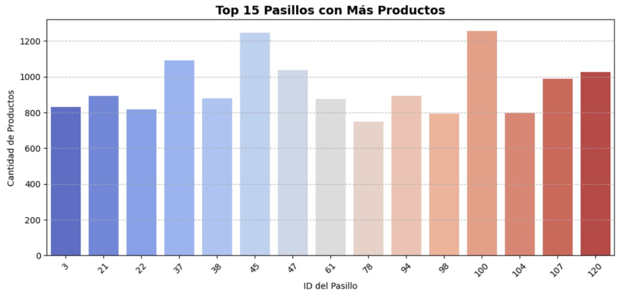


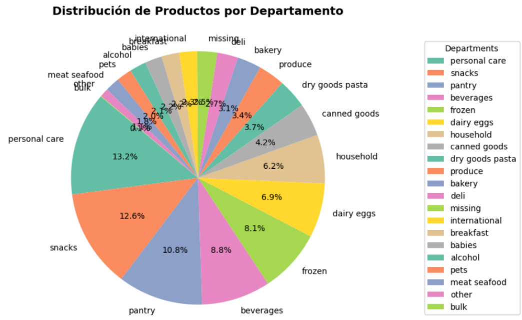




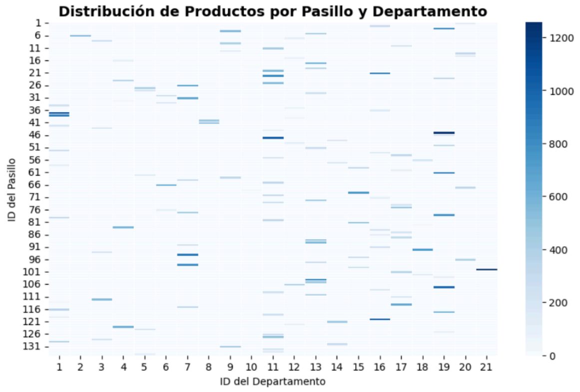


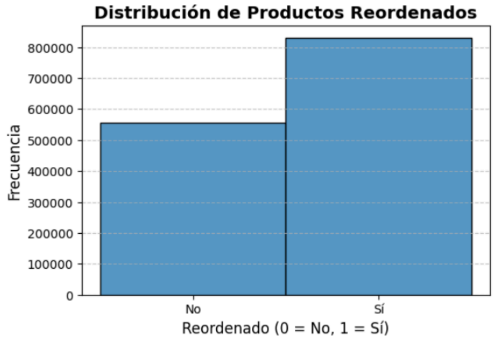

Cantidad de Productos por Departamento
Cantidad de Productos
ID del Departamento

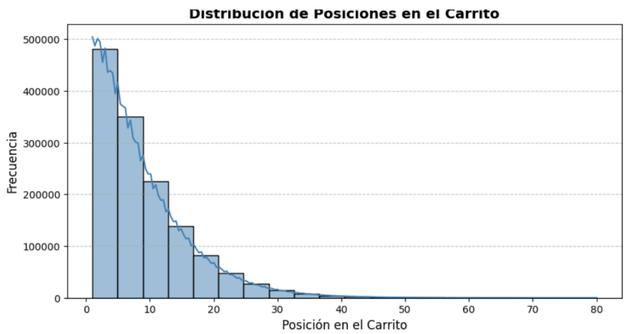

Top 15 Pasillos con Más Productos
Cantidad de Productos
ID del Pasillo

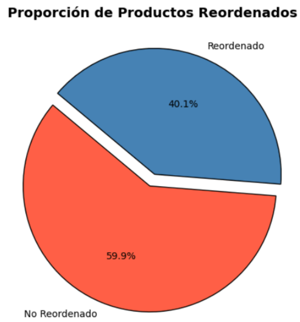

Distribución de Productos por Departamento
Departments
personal care
snacks
pantry
beverages
frozen
dairy eggs
household
canned goods
dry goods pasta
produce
bakery
deli
missing
international
breakfast
babies
alcohol
pets
meat seafood
other
bulk
13.2%
12.6%
10.8%
8.8%
8.1%
6.9%
6.2%
4.2%
3.7%
3.4%
3.1%

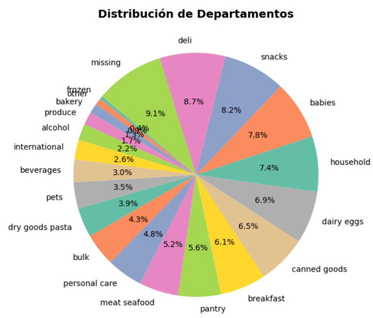

Palabras más comunes en los nombres de productos
Organic
Original
Gluten Free
Natural
Chicken
Ice Cream
Dark Chocolate
Classic
Strawberry
Sea Salt
Honey
Apple
Lemon
Peanut Butter
Vanilla
Premium
Sweet
Fresh
Bar
Cat Food
White
Pure
Light
Blend
Mini
Cheese


Distribución de Productos por Pasillo y Departamento
ID del Pasillo
ID del Departamento


Distribución de Productos Reordenados
Frecuencia
No
Sí
Reordenado (0 = No, 1 = Sí)


Distribución de Posiciones en el Carrito
Frecuencia
Posición en el Carrito


Proporción de Productos Reordenados
Reordenado
40.1%
59.9%
No Reordenado


Distribución de Departamentos
deli
snacks
babies
household
dairy eggs
canned goods
breakfast
pantry
meat seafood
personal care
bulk
dry goods pasta
pets
beverages
international
alcohol
produce
bakery
other
frozen
missing
9.1%
8.7%
8.2%
7.8%
7.4%
6.9%
6.5%
6.1%
5.6%
5.2%
4.8%
4.3%
3.9%
3.5%
3.0%
2.6%
2.2%
1.7%


Departamentos más comunes
frozen
goods
bakery
alcohol
international
produce
seafood
canned
meat
beverages
care
pets
bulk
pantry
personal
dairy
household
deli
babies
eggs
dry
pasta
snacks
breakfast
missing

• Análisis de correlaciones entre variables.

• **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.

    • Discusión de 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.