Instacart Market Basket Analysis Dataset

– Overview

The **Instacart Market Basket Analysis** dataset provides a detailed view of customer shopping behavior on Instacart, an online grocery delivery service. It includes millions of grocery orders made by users, along with product details, order history, and purchasing patterns. The dataset is structured to allow analysis of user shopping habits, product demand, and recommender system development.

- **Key Highlights:**
  - Contains **3 million+ grocery orders** from **200,000+ users**.
  - Covers **50,000+ products** across various **aisles and departments**.
  - Includes information on **reordering behavior**, **time of purchase**, and **product preferences**.
- **Dataset Structure:**

  The dataset consists of **six CSV files**, each providing different aspects of the order and product data:

  1. **orders.csv** – Details of customer orders, including order sequence, time of purchase, and user history.
  2. **order_products__prior.csv** – Product details for all past orders.
  3. **order_products__train.csv** – Product details for orders in the training dataset.
  4. **products.csv** – Product names and associated aisle and department IDs.
  5. **aisles.csv** – Mapping of aisle IDs to aisle names.
  6. **departments.csv** – Mapping of department IDs to department names.
- **Possible Use Cases:**
  - **Customer Behavior Analysis**: Understanding purchase frequency, reordering trends, and shopping habits.
  - **Market Basket Analysis**: Identifying frequently bought product combinations.
  - **Recommendation Systems**: Building personalized product recommendations.
  - **Retail Strategy Optimization**: Analyzing popular shopping hours, product demand, and category trends.
  - **Inventory Management**: Predicting demand for better stock management.

This dataset is widely used for machine learning applications, predictive analytics, and business intelligence.

1. **orders.csv:**

   This file contains information about customer orders placed on Instacart, an online grocery delivery service. Each row represents a single order, providing details about when the order was placed, the user who placed it, and whether it was their first order or a repeat purchase.

   Columns Description:

   - **order_id** (*integer*): A unique identifier for each order.
   - **user_id** (*integer*): A unique identifier for each customer.
   - **eval_set** (*string*): Specifies which dataset the order belongs to—either *"prior"*, *"train"*, or *"test"*.
   - **order_number** (*integer*): The sequential order count for the given user (e.g., 1 for their first order, 2 for their second, etc.).
   - **order_dow** (*integer*): The day of the week the order was placed (0 = Sunday, 6 = Saturday).
   - **order_hour_of_day** (*integer*): The hour of the day (0–23) when the order was placed.
   - **days_since_prior_order** (*float*): The number of days since the customer's previous order (NaN for first-time customers).

2. **products.csv:**

   The products.csv file contains information about the products available on Instacart. Each row represents a unique product, including its name, the department it belongs to, and the corresponding aisle. This data is crucial for analyzing product distribution, customer purchasing behavior, and basket composition.

   Columns Description:

   - **product_id** *(int)***:** A unique identifier for each product.
   - **product_name** *(string)*: The name of the product.
   - **aisle_id** *(int)*: A foreign key referencing the aisles.csv file, indicating the aisle where the product is located.
   - **department_id** *(int)*: A foreign key referencing the departments.csv file, indicating the broader category the product belongs to.

3. **order_products__prior.csv:**

   This file contains historical order details for repeat customers, helping to analyze customer behavior, product preferences, and shopping trends.

   Columns Description:

   - **order_id** *(int)*: Unique identifier for each order.
   - **product_id** *(int)*: Unique identifier for each product in the order.
   - **add_to_cart_order** *(int)*: The sequence in which the product was added to the cart.
   - **Reordered** *(int)*: A binary value (1 = product was previously ordered by the customer, 0 = first-time purchase).

4. **order_products__train.csv:**
   This file contains information about the products purchased in customer orders that are designated for model training.
   Columns Description:
   - **order_id** (int): Unique identifier for the order.
   - **product_id** (int): Unique identifier for the product purchased in the order.
   - **add_to_cart_order** (int): The sequence in which the product was added to the cart (e.g., first, second, third item).
   - **reordered** (int): A binary flag indicating whether the product was reordered (1 = reordered, 0 = first-time purchase).

5. **departments.csv:**
   It helps categorize products into broader groups, aiding in analysis and visualization of purchasing trends across different product categories.
   Columns Description:
   - **department_id** (Integer) – A unique identifier for each department.
   - **department** (String) – The name of the department (e.g., "produce," "dairy eggs," "beverages").

6. **aisles.csv:**
   It helps in organizing products into specific subcategories within broader departments.
   Columns Description:
   - **aisle_id** (Int): A unique identifier for each aisle.
   - **aisle** (String): The name of the aisle (e.g., "fresh fruits," "baking ingredients," "beverages")