Customer Filtering using DuckDB and Python

Overview

This project demonstrates customer data analysis using **DuckDB** and **Python**.
Two datasets were provided:

1. **customers.csv** — containing customer details (`id`, `name`, `age`, `email`)
2. **purchases.csv** — containing purchase records (`customer_id`, `product_id`, `quantity`)

The aim was to perform joins, grouping, filtering, and aggregation using SQL queries executed inside Python with DuckDB.

Tasks and Approach

1. Join and Group Data

- Joined `customers` and `purchases` tables on `customer_id`.
- Grouped results by customer ID, name, and age.
- Calculated the **total quantity of items purchased** by each customer.

Insight:
This step provided an aggregated view of how many items each customer has purchased over all transactions.

2. Filter by Quantity and Age

- From the grouped dataset, filtered customers:
    - Who purchased **more than 5 items** in total.

- Whose **age is less than 30**.

Insight:
This revealed high-purchasing younger customers, useful for targeted marketing campaigns.

---

3. Identify Customers with Multiple Product Categories

- Counted the number of **distinct products** each customer purchased.
- Filtered customers with **two or more distinct products**.

Insight:
These customers show diverse buying habits and may respond well to cross-selling offers.

4. Calculate Total Spending for Customers Aged 30+

- Selected customers aged **30 or older**.
- Summed the total quantity purchased.

Insight:
Older customers have different buying patterns; the aggregated spending figure helps in tailoring products and promotions for this group.

Tools Used

- **Python 3** for scripting
- **Pandas** for reading CSV files and handling DataFrames
- **DuckDB** for executing SQL queries directly on Pandas DataFrames

How to Run

1. Install dependencies in Colab:

```
pip install duckdb pandas
```

1. Upload `customers.csv` and `purchases.csv` to Colab.
2. Run the Python script provided in the project.
3. Results will be saved as:
    - `join_group.csv`
    - `filter.csv`
    - `multi_category.csv`
    - `total_spending.csv`

Key Observations

- Younger customers with high purchase volume tend to buy more frequently but fewer product categories.
- Some customers engage across multiple product categories, making them ideal for bundled promotions.
- Customers aged 30+ contribute a significant portion of total purchases, highlighting their long-term value.