Untitled

August 16, 2025

[1]:
```
%run /home/jovyan/project/hw-03/hw03_solution_jupyter.py
```

```
============================================================
HOMEWORK 03 - PySpark Data Analysis (Jupyter Version)
============================================================

 Task 1: Loading CSV files as DataFrames
----------------------------------------
 Users DataFrame loaded: 100 rows
root
 |-- user_id: integer (nullable = true)
 |-- name: string (nullable = true)
 |-- age: integer (nullable = true)
 |-- email: string (nullable = true)

+-------+------+---+-----------------+
|user_id|  name|age|            email|
+-------+------+---+-----------------+
|      1|User_1| 45|user1@example.com|
|      2|User_2| 48|user2@example.com|
|      3|User_3| 36|user3@example.com|
|      4|User_4| 46|user4@example.com|
|      5|User_5| 29|user5@example.com|
+-------+------+---+-----------------+
only showing top 5 rows

 Products DataFrame loaded: 50 rows
root
 |-- product_id: integer (nullable = true)
 |-- product_name: string (nullable = true)
 |-- category: string (nullable = true)
 |-- price: double (nullable = true)

+----------+------------+-----------+-----+
|product_id|product_name|   category|price|
+----------+------------+-----------+-----+
|         1|   Product_1|     Beauty|  8.3|
|         2|   Product_2|       Home|  8.3|
```

```
|         3|   Product_3|Electronics|  9.2|
|         4|   Product_4|Electronics|  2.6|
|         5|   Product_5|Electronics|  9.4|
+----------+------------+-----------+-----+
only showing top 5 rows

  Purchases DataFrame loaded: 200 rows
root
 |-- purchase_id: integer (nullable = true)
 |-- user_id: integer (nullable = true)
 |-- product_id: integer (nullable = true)
 |-- date: date (nullable = true)
 |-- quantity: integer (nullable = true)

+-----------+-------+----------+----------+--------+
|purchase_id|user_id|product_id|      date|quantity|
+-----------+-------+----------+----------+--------+
|          1|     52|         9|2022-01-01|       1|
|          2|     93|        37|2022-01-02|       8|
|          3|     15|        33|2022-01-03|       1|
|          4|     72|        42|2022-01-04|       9|
|          5|     61|        44|2022-01-05|       6|
+-----------+-------+----------+----------+--------+
only showing top 5 rows


  Data Loading Summary:
  • Users: 100 records
  • Products: 50 records
  • Purchases: 200 records

  Data Quality Check:
  • Users with null values: 2
  • Products with null values: 2
  • Purchases with null values: 3

  Task 1 completed successfully!
============================================================

  Task 2: Cleaning data by removing rows with missing values
----------------------------------------

Before cleaning:
  • Users with nulls: 2
  • Products with nulls: 3
  • Purchases with nulls: 5

  Users cleaned: 100 → 95 rows
```

```
  Products cleaned: 50 → 47 rows
  Purchases cleaned: 200 → 195 rows

  Cleaning Summary:
  • Users removed: 5 rows
  • Products removed: 3 rows
  • Purchases removed: 5 rows

  Task 3: Total purchase amount for each product category
----------------------------------------

Total purchase amount by category:
+-----------+------------------+
|   category|      total_amount|
+-----------+------------------+
|     Sports|1802.4999999999998|
|       Home|1523.4999999999998|
|Electronics|1174.7999999999997|
|   Clothing|             790.3|
|     Beauty| 459.8999999999999|
+-----------+------------------+


  Task 4: Purchase amount for age group 18-25
----------------------------------------

Users in age group 18-25: 20 users

Purchase amount by category for age 18-25:
+-----------+------------------+
|   category|total_amount_18_25|
+-----------+------------------+
|       Home|             361.1|
|     Sports|310.49999999999994|
|Electronics|             249.6|
|   Clothing|             245.0|
|     Beauty|41.400000000000006|
+-----------+------------------+


  Task 5: Share of purchases for age group 18-25
----------------------------------------

Total spending for age 18-25: $1207.60

Share of purchases by category for age 18-25:
+-----------+------------------+----------------+
|   category|total_amount_18_25|percentage_share|
```

```
+-----------+------------------+----------------+
|       Home|             361.1|            29.9|
|     Sports|310.49999999999994|           25.71|
|Electronics|             249.6|           20.67|
|   Clothing|             245.0|           20.29|
|     Beauty|41.400000000000006|            3.43|
+-----------+------------------+----------------+


  Task 6: Top 3 categories for age 18-25
--------------------------------------

Top 3 product categories with highest percentage for age 18-25:
+-----------+------------------+----------------+
|   category|total_amount_18_25|percentage_share|
+-----------+------------------+----------------+
|       Home|             361.1|            29.9|
|     Sports|310.49999999999994|           25.71|
|Electronics|             249.6|           20.67|
+-----------+------------------+----------------+


============================================================
  All homework tasks completed successfully!
============================================================
```