DWM PARTICAL NO 7

Assignment No 7: Demonstration of any ETL tool

sales.csv file

Link - https://github.com/Saikasote/dwm

Code:

```
import pandas as pd
import mysql.connector
from dateutil import parser

# === EXTRACT ===
df = pd.read_csv('sales_data.csv')

# === TRANSFORM ===

# 1. Clean column headers
df.columns = [col.lower().replace(" ", "_") for col in df.columns]

# 2. Normalize text fields (capitalize names, cities, etc.)
df['customer_name'] = df['customer_name'].str.strip().str.title()
df['city'] = df['city'].str.strip().str.title()
df['category'] = df['category'].str.strip().str.title()
df['product'] = df['product'].str.strip().str.title()

# 3. Standardize date format
def parse_date(date_str):
    try:
        return parser.parse(date_str).date()
    except:
        return None

df['date'] = df['date'].apply(parse_date)
df = df[df['date'].notnull()]  # Drop rows with invalid dates

# 4. Filter invalid quantities and prices
df = df[(df['quantity'] > 0) & (df['price'] > 0)]

# 5. Drop duplicates
df = df.drop_duplicates()

# 6. Add total_amount column
df['total_amount'] = df['quantity'] * df['price']
```

```
# === LOAD ===

# MySQL connection setup — update with your credentials
conn = mysql.connector.connect(
    host='*****',
    user='****',        # ← Change this
    password='*****',    # ← Change this
    database='*****'     # ← Change this
)
cursor = conn.cursor()

# Create table if it doesn't exist
cursor.execute("""
CREATE TABLE IF NOT EXISTS sales_data (
    id INT AUTO_INCREMENT PRIMARY KEY,
    customer_name VARCHAR(100),
    city VARCHAR(100),
    category VARCHAR(100),
    product VARCHAR(100),
    quantity INT,
    price FLOAT,
    date DATE,
    total_amount FLOAT
)
""")

# Insert data
for _, row in df.iterrows():
    cursor.execute("""
        INSERT INTO sales_data (customer_name, city, category, product, quantity, price, date,
total_amount)
        VALUES (%s, %s, %s, %s, %s, %s, %s, %s)
    """, (
        row['customer_name'], row['city'], row['category'], row['product'],
        int(row['quantity']), float(row['price']), row['date'], float(row['total_amount'])
    ))

conn.commit()
cursor.close()
conn.close()

print("Preprocessing and ETL completed. Clean data loaded into MySQL.")
```

Output:

1. *MySQL Table*

id	customer_name	city	category	product	quantity	price	date	total_amount
1	Alice	Bangalore	Home Appliances	Bed	5	5944	2023-09-18	29720
2	Vijay	Chennai	Books	Biography	2	46279	2023-06-01	92558
3	Charlie	Kolkata	Fashion	Shirt	1	42152	2023-03-05	42152
4	David	Bangalore	Electronics	Laptop	3	11797	2023-12-18	35391
5	Eva	Kolkata	Fashion	Shirt	2	26767	2023-10-17	53534
6	Jatin	Chennai	Home Appliances	Shirt	2	41431	2023-02-06	82862
7	Charlie	Chennai	Books	Fiction	4	22359	2023-08-17	89436
8	Vijay	Delhi	Home Appliances	Smartphone	3	13797	2023-03-11	41391
9	Vijay	Bangalore	Furniture	Chair	4	25878	2023-10-14	103512
10	Vijay	Kolkata	Grocery	Wheat	1	31408	2023-03-17	31408
11	Isha	Kolkata	Toys	Car	1	29130	2023-02-16	29130
12	Tony	Delhi	Electronics	Laptop	5	33382	2023-10-16	166910
13	Bob	Kolkata	Books	Textbook	1	33653	2023-06-22	33653
14	Isha	Delhi	Fashion	Shirt	1	27984	2023-04-15	27984
15	Isha	Chennai	Electronics	Laptop	4	4218	2023-11-24	16872
16	Grace	Kolkata	Home Appliances	Microwave	2	16970	2023-02-15	33940
17	Tony	Mumbai	Grocery	Sugar	4	11131	2023-03-15	44524
18	Grace	Mumbai	Furniture	Chair	3	30285	2023-08-31	90855
19	Isha	Bangalore	Fashion	Jacket	5	3421	2023-01-03	17105
20	Vijay	Delhi	Home Appliances	Textbook	2	22501	2023-02-15	45002
21	Tony	Bangalore	Grocery	Sugar	4	37425	2023-02-07	149700
22	Bob	Kolkata	Grocery	Wheat	1	24353	2023-03-27	24353
23	Vijay	Mumbai	Furniture	Chair	2	507	2023-07-12	1014

2. Python Script

```
Preprocessing and ETL completed. Clean data loaded into MySQL.
```