

COLLEGE OF COMPUTING

DEPARTMENT OF SOFTWARE ENGINEERING

FUNDAMENTAL OF

BIG DATA ANALYTICS AND BI

INDIVIDUAI ASSIGNMNT

**NAME ID**

**DESALEGN GEBEY** **DBU/040/14**

**SUBMITION DATE : 13/02/2025**

**SUBMITED TO : Derbew Felasman (MSc).**

### **Extracting and Analyzing Dataset using Pandas**

#### 1. **Importing Necessary Libraries**

import pandas as pd

from sqlalchemy import create\_engine

#### 2. **Loading the Dataset**

# Load the CSV file (Change the file path to match your actual file)

file\_path = r"C:\Users\Postlab\Desktop\Aliexpress\data\raw\_csv.csv"

# Read the CSV file into a pandas DataFrame

df = pd.read\_csv(file\_path)

#### 3. Displaying the First Five Rows

df.head()

#### Sample Output:

| **id** | **storeId** | **storeName** | **title** | **rating** |
| --- | --- | --- | --- | --- |
| 1005002508947205 | 900250272 | IDEACH Official Store | Wireless Handheld Vacuum Cleaner 10kPa 150W | 4.8 |
| 1005004546160959 | 5379102 | Factory Direct Collected Store | Portable Water Drop Humidifier USB Desktop | 4.8 |
| 1005004967489874 | 911794975 | HOMDD Specialty Store | Portable Desktop Instant Automatic Multi-Speed | 0.0 |
| 1005003601452827 | 912652146 | XINGHUO ONLINE Store | Portable 420ml Air Humidifier Aroma Oil Humidifier | 4.7 |
| 33022569739 | 4991203 | ABIR Official Store | ABIR X6 Robot Vacuum Cleaner Visual Navigation | 4.7 |

#### 4. Checking Dataset Shape (Rows and Columns)

df.shape

**Output:** (711424, 18)

#### 5. Displaying Column Names

df.columns

**Output:**

Index(['id', 'storeId', 'storeName', 'title', 'rating', 'launchTime',

'category', 'postCategory', 'sold', 'price', 'discount', 'shippingCost',

'imageUrl', 'storeUrl', 'category\_name', 'category\_id', 'type',

'total-revenue'],

dtype='object')

#### 6. Removing Duplicate Records

df = df.drop\_duplicates()

#### 7. Statistical Summary of the Dataset

df.describe()

#### 8. Calculating Total Revenue

df['total-revenue'] = df['price'] + df['shippingCost']

df

This document outlines the essential steps for extracting, cleaning, and analyzing the dataset using Pandas in Python.