

30-45 min each coding Assessment

Q1) You have been given a CSV file containing a list of user interactions on a website. Each row in the file represents a single interaction and contains the following fields:

CSV file = [user_visited.csv](#)

User ID

Timestamp

URL visited

Page view duration (in seconds)

Write a Python program that reads the CSV file and outputs the following information:

The total number of interactions in the file.

The total number of unique users in the file.

The most visited URL in the file.

The average time spent on each URL.

Your program should take the following inputs:

The path to the input CSV file

The path to the output file

Your program should write the output to a CSV file with the following fields:

Total Interactions

Total Unique Users

Most Visited URL

Average Time Spent on Each URL

You can assume that the CSV file is well-formed and contains no errors.

Some additional requirements:

The program should be able to handle large datasets efficiently.

The program should use the Pandas library to read and process the CSV file.

The program should be well-documented and include comments where necessary.

Q2: You have been given a CSV file that contains data about online purchases made by customers. The CSV file contains the following columns:

- transaction_id (int): Unique identifier for each transaction.
- customer_id (int): Unique identifier for each customer.
- timestamp (str): Timestamp of when the transaction was made in the format 'YYYY-MM-DD HH:MM:SS'.
- product_name (str): Name of the product purchased.
- product_category (str): Category of the product purchased.
- product_price (float): Price of the product purchased.
- quantity (int): Quantity of the product purchased.

You need to write a Python program that reads this [customer_input.csv](#) file, processes the data, and writes the output to a new CSV file. The output CSV file should contain the following columns:

- date (str): Date of the transaction in the format 'YYYY-MM-DD'.
- customer_id (int): Unique identifier for each customer.
- total_spent (float): Total amount spent by the customer on that date.

Your program should perform the following tasks:

1. Read the input CSV file and parse the data.
2. Group the data by date and customer_id.
3. Calculate the total amount spent by each customer on each date.
4. Write the output CSV file with the specified columns and format.

You may assume that the input CSV file is properly formatted and that all columns are present and correctly typed.

Example Input CSV File:

```
transaction_id,customer_id,timestamp,product_name,product_category,product_price,quantity
1,1,2022-03-01 10:00:00,Product A,Category 1,10.00,1
2,2,2022-03-01 11:00:00,Product B,Category 2,20.00,2
3,1,2022-03-01 12:00:00,Product C,Category 1,15.00,3
4,3,2022-03-02 09:00:00,Product D,Category 2,30.00,1
5,1,2022-03-02 11:00:00,Product E,Category 1,25.00,2
```

Example Output CSV File:

```
date,customer_id,total_spent
2022-03-01,1,45.00
2022-03-01,2,40.00
2022-03-02,1,50.00
2022-03-02,3,30.00
```

Situation based question:

Q3: You are tasked with designing a data pipeline for a new e-commerce website. The website generates a large amount of data on customer purchases, including information on the products purchased, the customer's location, and the time of the purchase. Your goal is to design a pipeline that can efficiently process this data and provide insights to the business team. The pipeline should include the following steps:

1. Ingest the raw data from the website's database.
2. Clean and transform the data as necessary.
3. Store the processed data in a data warehouse.
4. Run regular analytics queries on the data to provide insights to the business team.

Describe the key components of your proposed data pipeline, including the technologies you would use and any challenges you anticipate. How would you ensure that the pipeline is reliable, scalable, and efficient?